

Bregman Voronoi Diagrams: Properties, Algorithms and Applications*

Frank Nielsen[†] Jean-Daniel Boissonnat[‡] Richard Nock[§]

Abstract

The Voronoi diagram of a finite set of objects is a fundamental geometric structure that subdivides the embedding space into regions, each region consisting of the points that are closer to a given object than to the others. We may define many variants of Voronoi diagrams depending on the class of objects, the distance functions and the embedding space. In this paper, we investigate a framework for defining and building Voronoi diagrams for a broad class of distance functions called Bregman divergences. Bregman divergences include not only the traditional (squared) Euclidean distance but also various divergence measures based on entropic functions. Accordingly, Bregman Voronoi diagrams allow to define information-theoretic Voronoi diagrams in statistical parametric spaces based on the relative entropy of distributions. We define several types of Bregman diagrams, establish correspondences between those diagrams (using the Legendre transformation), and show how to compute them efficiently. We also introduce extensions of these diagrams, e.g. k -order and k -bag Bregman Voronoi diagrams, and introduce Bregman triangulations of a set of points and their connexion with Bregman Voronoi diagrams. We show that these triangulations capture many of the properties of the celebrated Delaunay triangulation. Finally, we give some applications of Bregman Voronoi diagrams which are of interest in the context of computational geometry and machine learning.

Categories and Subject Descriptors: I.3.5 [Computer Graphics] Computational Geometry and Object Modeling — Geometric algorithms, languages, and systems; F.2.2 [Analysis of Algorithms and Problem Complexity]: Nonnumerical Algorithms and Problems — Geometrical problems and computations; G.2.1 [Discrete Mathematics]: Combinatorics.

*A preliminary version appeared in the 18th ACM-SIAM Symposium on Discrete Algorithms (SODA), pp. 746-755, 2007. Related materials are available online at <http://www.csl.sony.co.jp/person/nielsen/BregmanVoronoi/> — Version for arXiv (small resolution jpg/png files).

[†]Sony Computer Science Laboratories Inc., Fundamental Research Laboratory, Japan.

[‡]INRIA Sophia-Antipolis, GEOMETRICA, France.

[§]Université Antilles-Guyane, CEREGMIA, France.

General Terms: Algorithms, Theory

Keywords: Computational Information Geometry, Voronoi diagram, Delaunay triangulation, Bregman divergence, Bregman ball, Legendre transformation, Quantification, Sampling, Clustering

1 Introduction and prior work

The *Voronoi diagram* $\text{vor}(\mathcal{S})$ of a set of n points $\mathcal{S} = \{\mathbf{p}_1, \dots, \mathbf{p}_n\}$ of the d -dimensional Euclidean space \mathbb{R}^d is defined as the *cell complex* whose d -cells are the *Voronoi regions* $\{\text{vor}(\mathbf{p}_i)\}_{i \in \{1, \dots, n\}}$ where $\text{vor}(\mathbf{p}_i)$ is the set of points of \mathbb{R}^d closer to \mathbf{p}_i than to any other point of \mathcal{S} with respect to a *distance function* δ :

$$\text{vor}(\mathbf{p}_i) \stackrel{\text{def}}{=} \{\mathbf{x} \in \mathbb{R}^d \mid \delta(\mathbf{p}_i, \mathbf{x}) \leq \delta(\mathbf{p}_j, \mathbf{x}) \ \forall \ \mathbf{p}_j \in \mathcal{S}\}.$$

Points $\{\mathbf{p}_i\}_i$ are called the *Voronoi sites* or *Voronoi generators*. Since its inception in disguise by Descartes in the 17th century [5], Voronoi diagrams have found a broad spectrum of applications in science. Computational geometers have focused at first on *Euclidean* Voronoi diagrams [5] by considering the case where $\delta(\mathbf{x}, \mathbf{y})$ is the Euclidean distance $\|\mathbf{x} - \mathbf{y}\| = \sqrt{\sum_{i=1}^d (x_i - y_i)^2}$. Voronoi diagrams have been later on defined and studied for other distance functions, most notably the L_1 distance $\|\mathbf{x} - \mathbf{y}\|_1 = \sum_{i=1}^d |x_i - y_i|$ (Manhattan distance) and the L_∞ distance $\|\mathbf{x} - \mathbf{y}\|_\infty = \max_{i \in \{1, \dots, d\}} |x_i - y_i|$ [10, 5]. Klein further presented an *abstract framework* for describing and computing the fundamental structures of abstract Voronoi diagrams [26, 11].

In artificial intelligence, machine learning techniques also rely on geometric concepts for *building classifiers* in supervised problems (*e.g.*, linear separators, oblique decision trees, etc.) or *clustering data* in unsupervised settings (*e.g.*, k -means, support vector clustering [2], etc.). However, the considered data sets \mathcal{S} and their underlying spaces \mathcal{X} are usually *not* metric spaces. The notion of distance between two elements of \mathcal{X} needs to be replaced by a *pseudo-distance* that is not necessarily symmetric and may not satisfy the *triangle inequality*. Such a pseudo-distance is also referred to as *distortion*, *(dis)similarity* or *divergence* in the literature. For example, in parametric statistical spaces \mathcal{X} , a vector point represent a distribution and its coordinates store the parameters of the associated distribution. A notion of “distance” between two such points is then needed to represent the divergence between the corresponding distributions.

Very few works have tackled an in-depth study of Voronoi diagrams and their applications for such a kind of statistical spaces. This is all the more important even for ordinary Voronoi diagrams as Euclidean point location of sites are usually *observed* in *noisy* environments (*e.g.*, imprecise point measures in computer vision experiments), and “noise” is often modeled by means of Normal distributions (so-called “Gaussian noise”). To the best of our knowledge, statistical Voronoi diagrams have only been considered in a 4-page short paper of Onishi and

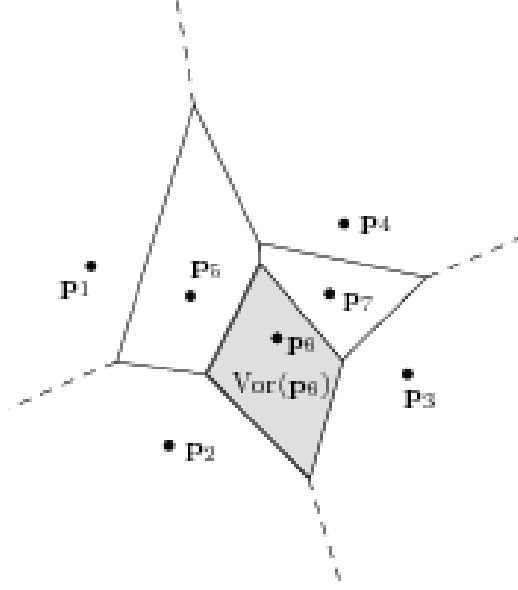


Figure 1: Ordinary Euclidean Voronoi diagram of a given set \mathcal{S} of seven sites. In the bounded Voronoi cell $\text{vor}(\mathbf{p}_6)$, every point $\mathbf{p} \in \text{vor}(\mathbf{p}_6)$ is closer to \mathbf{p}_6 than to any other site of \mathcal{S} (with respect to the Euclidean distance). Dashed segments denote infinite edges delimiting unbounded cells.

Imai [34] which relies on Kullback-Leibler divergence of dD multivariate normal distributions to study combinatorics of their Voronoi diagrams, and subsequently in a 2-page video paper of Sadakane et al. [40] which defines the divergence implied by a convex function and its conjugate, and present the Voronoi diagram with flavors of information geometry [1] (see also [35] and related short communications [25, 24]). Our study of Bregman Voronoi diagrams generalizes and subsumes these preliminary studies using an easier concept of divergence: Bregman divergences [12, 6] that do not rely *explicitly* on convex conjugates. Bregman divergences encapsulate the squared Euclidean distance and many widely used divergences, e.g. the Kullback-Leibler divergence. It should be noticed however that other divergences have been defined and studied in the context of Riemannian geometry [1]. Sacrificing for some generality, while not very restrictive in practice, allows a much simpler treatment and our study of Bregman divergences is elementary and does not rely on Riemannian geometry.

In this paper, we give a thorough treatment of Bregman Voronoi diagrams which elegantly *unifies* the ordinary Euclidean Voronoi diagram and statistical Voronoi diagrams. Our contributions are summarized as follows:

- Since Bregman divergences are not symmetric, we define *two types* of Bregman Voronoi diagrams. One is an affine diagram with convex polyhedral cells while the other one is curved. The cells of those two diagrams are in 1-1 correspondence through the Legendre transformation. We also introduce a *third-type* symmetrized Bregman Voronoi diagram.

- We present a simple way to compute the Bregman Voronoi diagram of a set of points by lifting the points in a higher dimensional space using an extra dimension. This mapping leads also to combinatorial bounds on the size of these diagrams. We also define weighted Bregman Voronoi diagrams and show that the class of these diagrams is identical to the class of affine (or power) diagrams. Special cases of weighted Bregman Voronoi diagrams are the k -order and k -bag Bregman Voronoi diagrams.
- We define two triangulations of a set of points. The first one captures some of the most important properties of the well-known Delaunay triangulation. The second triangulation is called a geodesic Bregman triangulation since its edges are geodesic arcs. Differently from the first triangulation, this triangulation is the geometric dual of the first-type Bregman Voronoi diagram of its vertices.
- We give a few applications of Bregman Voronoi diagrams which are of interest in the context of computational geometry and machine learning.

The outline of the paper is as follows: In Section 2, we define Bregman divergences and recall some of their basic properties. In Section 3, we study the geometry of Bregman spaces and characterize bisectors, balls and geodesics. Section 4 is devoted to Bregman Voronoi diagrams and Section 5 to Bregman triangulations. In Section 6, we select a few applications of interest in computational geometry and machine learning. Finally, Section 7 concludes the paper and mentions further ongoing investigations.

Notations. In the whole paper, \mathcal{X} denotes an open convex domain of \mathbb{R}^d and $F : \mathcal{X} \mapsto \mathbb{R}$ a strictly convex and differentiable function. \mathcal{F} denotes the graph of F , i.e. the set of points $(\mathbf{x}, z) \in \mathcal{X} \times \mathbb{R}$ where $z = F(\mathbf{x})$. We write $\hat{\mathbf{x}}$ for the point $(\mathbf{x}, F(\mathbf{x})) \in \mathcal{F}$. ∇F , $\nabla^2 F$ and $\nabla^{-1} F$ denote respectively the gradient, the Hessian and the inverse gradient of F .

2 Bregman divergences

In this section, we recall the definition of Bregman¹ divergences and some of their main properties (§2.1). We show that the notion of Bregman divergence encapsulates the squared Euclidean distance as well as several well-known information-theoretic divergences. We introduce the notion of dual divergences (§2.2) and show how this comes in handy for symmetrizing Bregman divergences (§2.3). Finally, we prove that the Kullback-Leibler divergence of distributions that belong to the exponential family of distributions can be viewed as a Bregman divergence (§2.4).

¹Lev M. Bregman historically pioneered this notion in the seminal work [12] on minimization of a convex objective function under linear constraints. See <http://www.math.bgu.ac.il/serv/segel/bregman.html>. We gratefully acknowledge him for sending us this historical paper.

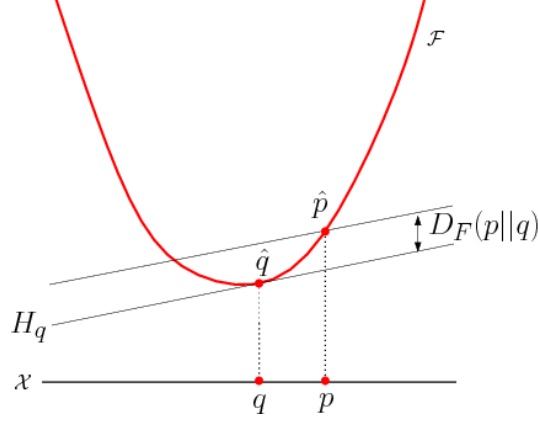


Figure 2: Visualizing the Bregman divergence. $D_F(\cdot||\mathbf{q})$ is the vertical distance between \mathcal{F} and the hyperplane tangent to \mathcal{F} at $\hat{\mathbf{q}}$.

2.1 Definition and basic properties

For any two points \mathbf{p} and \mathbf{q} of $\mathcal{X} \subseteq \mathbb{R}^d$, the Bregman divergence² $D_F(\cdot||\cdot) : \mathcal{X} \mapsto \mathbb{R}$ of \mathbf{p} to \mathbf{q} associated to a strictly convex and differentiable function F (called the *generator function* of the divergence) is defined as

$$D_F(\mathbf{p}||\mathbf{q}) \stackrel{\text{def}}{=} F(\mathbf{p}) - F(\mathbf{q}) - \langle \nabla F(\mathbf{q}), \mathbf{p} - \mathbf{q} \rangle, \quad (1)$$

where $\nabla F = [\frac{\partial F}{\partial x_1} \dots \frac{\partial F}{\partial x_d}]^T$ denotes the gradient operator, and $\langle \mathbf{p}, \mathbf{q} \rangle$ the inner (or dot) product: $\sum_{i=1}^d p_i q_i$.

Informally speaking, Bregman divergence D_F is the *tail* of the Taylor expansion of F . See [16] for an axiomatic characterization of Bregman divergences as “permissible” divergences.

Lemma 1 *The Bregman divergence $D_F(\mathbf{p}||\mathbf{q})$ is geometrically measured as the vertical distance between $\hat{\mathbf{p}}$ and the hyperplane $H_{\mathbf{q}}$ tangent to \mathcal{F} at point $\hat{\mathbf{q}}$: $D_F(\mathbf{p}||\mathbf{q}) = F(\mathbf{p}) - H_{\mathbf{q}}(\mathbf{p})$.*

Proof: The tangent hyperplane to hypersurface $\mathcal{F} : z = F(\mathbf{x})$ at point $\hat{\mathbf{q}}$ is $H_{\mathbf{q}} : z = F(\mathbf{q}) + \langle \nabla F(\mathbf{q}), \mathbf{x} - \mathbf{q} \rangle$. It follows that $D_F(\mathbf{p}||\mathbf{q}) = F(\mathbf{p}) - H_{\mathbf{q}}(\mathbf{p})$ (see Figure 2). \square

We now give some basic properties of Bregman divergences. The first property seems to be new. The others are well known. First, observe that, for most functions F , the associated Bregman divergence is *not* symmetric, i.e. $D_F(\mathbf{p}||\mathbf{q}) \neq D_F(\mathbf{q}||\mathbf{p})$ (the symbol $||$ is put to emphasize this point, as is standard in information theory). The following lemma proves this claim.

²See JavaTM applet at <http://www.cs1.sony.co.jp/person/nielsen/BregmanDivergence/>

Lemma 2 *Let F be properly defined for D_F to exist. Then D_F is symmetric if and only if the Hessian $\nabla^2 F$ is constant on \mathcal{X} .*

Proof: (\Rightarrow) From Eq. 1, the symmetry $D_F(\mathbf{p}||\mathbf{q}) = D_F(\mathbf{q}||\mathbf{p})$ yields:

$$F(\mathbf{p}) = F(\mathbf{q}) + \frac{1}{2}\langle \mathbf{p} - \mathbf{q}, \nabla F(\mathbf{q}) + \nabla F(\mathbf{p}) \rangle. \quad (2)$$

A Taylor expansion of F around \mathbf{q} using the Lagrange form of the remainder also yields:

$$F(\mathbf{p}) = F(\mathbf{q}) + \langle \mathbf{p} - \mathbf{q}, \nabla F(\mathbf{q}) \rangle + \frac{1}{2}(\mathbf{p} - \mathbf{q})^T \nabla^2 F(\mathbf{q})(\mathbf{p} - \mathbf{q}) + \frac{1}{6}\langle \mathbf{p} - \mathbf{q}, \nabla F \rangle^3(\mathbf{r}_{\mathbf{pq}}) \quad (3)$$

with $\mathbf{r}_{\mathbf{pq}}$ on the line segment \mathbf{pq} . Equations (2) and (3) yield the following constraint:

$$\langle \mathbf{p} - \mathbf{q}, \nabla F(\mathbf{p}) \rangle = \langle \mathbf{p} - \mathbf{q}, \nabla F(\mathbf{q}) \rangle + (\mathbf{p} - \mathbf{q})^T \nabla^2 F(\mathbf{q})(\mathbf{p} - \mathbf{q}) + \frac{1}{3}\langle \mathbf{p} - \mathbf{q}, \nabla F \rangle^3(\mathbf{r}_{\mathbf{pq}}) \quad (4)$$

On the other hand, if we make the Taylor expansion of ∇F around \mathbf{q} and then multiply both sides by $\mathbf{p} - \mathbf{q}$, we separately obtain:

$$\langle \mathbf{p} - \mathbf{q}, \nabla F(\mathbf{p}) \rangle = \langle \mathbf{p} - \mathbf{q}, \nabla F(\mathbf{q}) \rangle + (\mathbf{p} - \mathbf{q})^T \nabla^2 F(\mathbf{q})(\mathbf{p} - \mathbf{q}) + \frac{1}{2}\langle \mathbf{p} - \mathbf{q}, \nabla F \rangle^3(\mathbf{s}_{\mathbf{pq}}),$$

with $\mathbf{s}_{\mathbf{pq}}$ on the line segment \mathbf{pq} . However, for this to equal Eq. (4), we must have $\langle \mathbf{p} - \mathbf{q}, \nabla F \rangle^3(\mathbf{r}_{\mathbf{pq}}) = (3/2)\langle \mathbf{p} - \mathbf{q}, \nabla F \rangle^3(\mathbf{s}_{\mathbf{pq}})$ for each \mathbf{p} and \mathbf{q} in \mathcal{X} . If we pick \mathbf{p} and \mathbf{q} very close to each other, this equality cannot be true, except when the third differentials are all zero on $\mathbf{r}_{\mathbf{pq}}$ and $\mathbf{s}_{\mathbf{pq}}$. Repeating this argument over each subset of \mathcal{X} having non zero measure, we obtain that the third differentials of F must be zero everywhere but on subsets of \mathcal{X} with zero measure, which implies that the second differentials (the Hessian of F , $\nabla^2 F$) are *constant* everywhere on \mathcal{X} .

(\Leftarrow) Assume the hessian $\nabla^2 F$ is constant on \mathcal{X} . In this case, because F is strictly convex, the Hessian $\nabla^2 F$ is positive definite, and we can factor it as $\nabla^2 F = \mathbf{P}^{-1}\mathbf{D}\mathbf{P}$ where \mathbf{D} is a diagonal matrix and \mathbf{P} a unitary rotation matrix. Reasoning in the basis of \mathcal{X} formed by \mathbf{P} , each element \mathbf{x} is mapped to $\mathbf{P}\mathbf{x}$, and we have $F(\mathbf{x}) = \sum_i d_i x_i^2$, where the d_i 's are the diagonal coefficients of \mathbf{D} . The symmetry of D_F is then immediate (i.e., D_F is a generalized quadratic distance). \square

Property 1 (Non-negativity) *The strict convexity of generator function F implies that, for any \mathbf{p} and \mathbf{q} in \mathcal{X} , $D_F(\mathbf{p}||\mathbf{q}) \geq 0$, with $D_F(\mathbf{p}||\mathbf{q}) = 0$ if and only if $\mathbf{p} = \mathbf{q}$.*

Property 2 (Convexity) *Function $D_F(\mathbf{p}||\mathbf{q})$ is convex in its first argument \mathbf{p} but not necessarily in its second argument \mathbf{q} .*

Bregman divergences can easily be constructed from simpler ones. For instance, multivariate Bregman divergences D_F can be created from univariate generator functions coordinate-wise as $F(\mathbf{x}) = \sum_{i=1}^d f_i(x_i)$ with $\nabla F = [\frac{df_1}{dx_1} \dots \frac{df_d}{dx_d}]^T$.

Because positive linear combinations of strictly convex and differentiable functions are strictly convex and differentiable functions, new generator functions (and corresponding Bregman divergences) can also be built as positive linear combinations of elementary generator functions. This is an important property as it allows to handle mixed data sets of heterogenous types in a unified framework.

Property 3 (Linearity) *Bregman divergence is a linear operator, i.e., for any two strictly convex and differentiable functions F_1 and F_2 defined on \mathcal{X} and for any $\lambda \geq 0$:*

$$D_{F_1 + \lambda F_2}(\mathbf{p}||\mathbf{q}) = D_{F_1}(\mathbf{p}||\mathbf{q}) + \lambda D_{F_2}(\mathbf{p}||\mathbf{q}).$$

Property 4 (Invariance under linear transforms) *$G(\mathbf{x}) = F(\mathbf{x}) + \langle \mathbf{a}, \mathbf{x} \rangle + b$, with $\mathbf{a} \in \mathbb{R}^d$ and $b \in \mathbb{R}$, is a strictly convex and differentiable function on \mathcal{X} , and $D_G(\mathbf{p}||\mathbf{q}) = D_F(\mathbf{p}||\mathbf{q})$.*

Examples of Bregman divergences are the squared Euclidean distance (obtained for $F(\mathbf{x}) = \|\mathbf{x}\|^2$ and the generalized quadratic distance function $F(\mathbf{x}) = \mathbf{x}^T \mathbf{Q} \mathbf{x}$ where \mathbf{Q} is a positive definite matrix. When \mathbf{Q} is taken to be the inverse of the variance-covariance matrix, D_F is the Mahalanobis distance, extensively used in computer vision. More importantly, the notion of Bregman divergence encapsulates various information measures based on entropic functions such as the Kullback-Leibler divergence based on the (unnormalized) Shannon entropy, or the Itakura-Saito divergence based on Burg entropy (commonly used in sound processing). Table 1 lists the main univariate Bregman divergences.

2.2 Legendre duality

We now turn to an essential notion of convex analysis: Legendre transform that will allow us to associate to any Bregman divergence a dual Bregman divergence.

Let F be a strictly convex and differentiable real-valued function on \mathcal{X} . The Legendre transformation makes use of the duality relationship between points and lines to associate to F a *convex conjugate* function $F^* : \mathbb{R}^d \mapsto \mathbb{R}$ given by [38]:

$$F^*(\mathbf{y}) = \sup_{\mathbf{x} \in \mathcal{X}} \{\langle \mathbf{y}, \mathbf{x} \rangle - F(\mathbf{x})\}.$$

The supremum is reached at the *unique* point where the gradient of $G(\mathbf{x}) = \langle \mathbf{y}, \mathbf{x} \rangle - F(\mathbf{x})$ vanishes or, equivalently, when $\mathbf{y} = \nabla F(\mathbf{x})$.

Dom. \mathcal{X}	Function F	Gradient	Inv. grad.	Divergence $D_F(p q)$
\mathbb{R}	Squared function x^2	$2x$	$\frac{x}{2}$	Squared loss (norm) $(p - q)^2$
$\mathbb{R}_+, \alpha \in \mathbb{N}$ $\alpha > 1$	Norm-like x^α	$\alpha x^{\alpha-1}$	$(\frac{x}{\alpha})^{\frac{1}{\alpha-1}}$	Norm-like $p^\alpha + (\alpha - 1)q^\alpha - \alpha p q^{\alpha-1}$
\mathbb{R}^+	Unnorm. Shannon entropy $x \log x - x$	$\log x$	$\exp(x)$	Kullback-Leibler div. (I-div.) $p \log \frac{p}{q} - p + q$
\mathbb{R}	Exponential $\exp x$	$\exp x$	$\log x$	Exponential loss $\exp(p) - (p - q + 1) \exp(q)$
\mathbb{R}^{+*}	Burg entropy $-\log x$	$-\frac{1}{x}$	$-\frac{1}{x}$	Itakura-Saito divergence $\frac{p}{q} - \log \frac{p}{q} - 1$
$[0, 1]$	Bit entropy $x \log x + (1 - x) \log(1 - x)$	$\log \frac{x}{1-x}$	$\frac{\exp x}{1+\exp x}$	Logistic loss $p \log \frac{p}{q} + (1 - p) \log \frac{1-p}{1-q}$
\mathbb{R}	Dual bit entropy $\log(1 + \exp x)$	$\frac{\exp x}{1+\exp x}$	$\log \frac{x}{1-x}$	Dual logistic loss $\log \frac{1+\exp p}{1+\exp q} - (p - q) \frac{\exp q}{1+\exp q}$
$[-1, 1]$	Hellinger-like $-\sqrt{1 - x^2}$	$\frac{x}{\sqrt{1-x^2}}$	$\frac{x}{\sqrt{1+x^2}}$	Hellinger-like $\frac{1-pq}{\sqrt{1-q^2}} - \sqrt{1-p^2}$

Table 1: Some common univariate Bregman divergences D_F .

As is well-known, F^* is strictly convex. To see this, consider the epigraph $\text{epi}(F^*)$, i.e. the set of points (\mathbf{y}, z) such that $F^*(\mathbf{y}) \leq z$. Clearly, $(\mathbf{y}, z) \in \text{epi}(F^*)$ iff $G_{\mathbf{x}}(\mathbf{y}) = \langle \mathbf{y}, \mathbf{x} \rangle - F(\mathbf{x}) \leq z$ for all $\mathbf{x} \in \mathcal{X}$. Therefore, $\text{epi}(F^*) = \cap_{\mathbf{x} \in \mathcal{X}} \text{epi}(G_{\mathbf{x}})$. Since $G_{\mathbf{x}}(\mathbf{y})$ is an affine function, $\text{epi}(G_{\mathbf{x}})$ is a half-space and $\text{epi}(F^*)$ being the intersection of half-spaces is a convex set, which proves that F^* is convex. The strict convexity follows from the fact that otherwise, F would not be differentiable in at least one point $\mathbf{z} \in \mathcal{X}$: at this point, $\langle \mathbf{y}_\alpha, \mathbf{z} \rangle - F(\mathbf{z}) \geq \langle \mathbf{y}_\alpha, \mathbf{x} \rangle - F(\mathbf{x}), \forall \mathbf{x} \in \mathcal{X}$, and $\mathbf{y}_\alpha = \alpha \mathbf{y}_1 + (1 - \alpha) \mathbf{y}_2, \forall \alpha \in [0, 1]$, $\mathbf{y}_1 \mathbf{y}_2$ being a segment on which F^* is not strictly convex. Thus, $\mathbf{y}_1 \mathbf{y}_2$ would be a subdifferential of F in \mathbf{z} contradicting the fact that F is differentiable.

For convenience, we write $\mathbf{x}' = \nabla F(\mathbf{x})$ (omitting the F in the \mathbf{x}' notation as it should be clear from the context). Figure 3 gives a geometric interpretation of the Legendre transformation. Using this notation, Eq. 1 can be rewritten as

$$D_F(\mathbf{p}||\mathbf{q}) = F(\mathbf{p}) - F(\mathbf{q}) - \langle \mathbf{q}', \mathbf{p} - \mathbf{q} \rangle. \quad (5)$$

Since F is a strictly convex and differentiable real-valued function on \mathcal{X} , its gradient ∇F is well defined as well as its inverse $\nabla^{-1} F$. Writing \mathcal{X}' for the *gradient space* $\{\nabla F(\mathbf{x}) = \mathbf{x}' | \mathbf{x} \in \mathcal{X}\}$, the convex conjugate F^* of F is the function: $\mathcal{X}' \subset \mathbb{R}^d \mapsto \mathbb{R}$ defined by

$$F^*(\mathbf{x}') = \langle \mathbf{x}, \mathbf{x}' \rangle - F(\mathbf{x}). \quad (6)$$

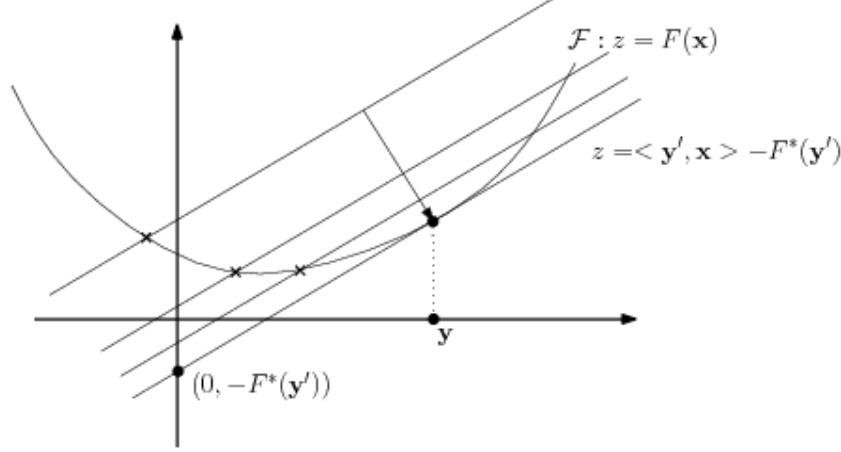


Figure 3: Legendre transformation of a strictly convex function F : The z -intercept $(0, -F^*(\mathbf{y}'))$ of the tangent hyperplane $H_{\mathbf{y}} : z = \langle \mathbf{y}', \mathbf{x} \rangle - F^*(\mathbf{y}')$ of \mathcal{F} at $\hat{\mathbf{y}}$ defines the value of the Legendre transform F^* for the dual coordinate $\mathbf{y}' = \nabla F(\mathbf{y})$. Any hyperplane passing through an other point of \mathcal{F} and parallel to $H_{\mathbf{y}}$ necessarily intersects the z -axis above $-F^*(\mathbf{y}')$.

Deriving this expression, we get

$$\langle \nabla F^*(\mathbf{x}'), d\mathbf{x}' \rangle = \langle \mathbf{x}, d\mathbf{x}' \rangle + \langle \mathbf{x}', d\mathbf{x} \rangle - \langle \nabla F(\mathbf{x}), d\mathbf{x} \rangle = \langle \mathbf{x}, d\mathbf{x}' \rangle = \langle \nabla^{-1} F(\mathbf{x}'), d\mathbf{x}' \rangle,$$

from which we deduce that $\nabla F^* = \nabla^{-1} F$. From Eq. 6, we also deduce $(F^*)^* = F$.

From the above discussion, it follows that D_{F^*} is a Bregman divergence, which we call the *Legendre dual divergence* of D_F . We have :

Lemma 3 $D_F(\mathbf{p}||\mathbf{q}) = F(\mathbf{p}) + F^*(\mathbf{q}') - \langle \mathbf{p}, \mathbf{q}' \rangle = D_{F^*}(\mathbf{q}'||\mathbf{p}')$

Proof: By Eq. 5, $D_F(\mathbf{p}||\mathbf{q}) = F(\mathbf{p}) - F(\mathbf{q}) - \langle \mathbf{p} - \mathbf{q}, \mathbf{q}' \rangle$, and, according to Eq. 6, we have $F(\mathbf{p}) = \langle \mathbf{p}', \mathbf{p} \rangle - F^*(\mathbf{p}')$ and $F(\mathbf{q}) = \langle \mathbf{q}', \mathbf{q} \rangle - F^*(\mathbf{q}')$. Hence, $D_F(\mathbf{p}||\mathbf{q}) = \langle \mathbf{p}', \mathbf{p} \rangle - F^*(\mathbf{p}') - \langle \mathbf{p}, \mathbf{q}' \rangle + F^*(\mathbf{q}') = D_{F^*}(\mathbf{q}'||\mathbf{p}')$ since $\mathbf{p} = \nabla F^{-1} \nabla F(\mathbf{p}) = \nabla F^*(\mathbf{p}')$. \square

Observe that, when D_F is symmetric, D_{F^*} is also symmetric.

The Legendre transform of the quadratic form $F(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T \mathbf{Q} \mathbf{x}$, where \mathbf{Q} is a symmetric invertible matrix, is $F^*(\mathbf{y}) = \frac{1}{2} \mathbf{y}^T \mathbf{Q}^{-1} \mathbf{y}$ (corresponding divergences D_F and D_{F^*} are both generalized quadratic distances).

To compute F^* , we use the fact that $\nabla F^* = \nabla^{-1} F$ and obtain F^* as $F^* = \int \nabla^{-1} F$. For example, the Hellinger-like measure is obtained by setting $F(x) = -\sqrt{1-x^2}$ (see Table 1). The inverse gradient is $\frac{x}{\sqrt{1+x^2}}$ and the dual convex conjugate is $\int \frac{x dx}{\sqrt{1+x^2}} = \sqrt{1+x^2}$. Integrating functions symbolically may be difficult or even not possible, and, in some cases, it will be required to approximate numerically the inverse gradient $\nabla^{-1} F(\mathbf{x})$.

Let us consider the univariate generator functions defining the divergences of Table 1. Both the squared function $F(\mathbf{x}) = x^2$ and Burg entropy $F(x) = -\log x$ are *self-dual*, i.e. $F = F^*$. This is easily seen by noticing that the gradient and inverse gradient are identical (up to some constant factor).

For the exponential function $F(x) = \exp x$, we have $F^*(y) = y \log y - y$ (the unnormalized Shannon entropy) and for the dual bit entropy $F(x) = \log(1 + \exp x)$, we have $F^*(y) = y \log \frac{y}{1-y} + \log(1 - y)$, the bit entropy. Note that the bit entropy function is a particular Bregman generator satisfying $F(x) = F(1 - x)$.

2.3 Symmetrized Bregman divergences

For non-symmetric d -variate Bregman divergences D_F , we define the *symmetrized divergence*

$$S_F(\mathbf{p}, \mathbf{q}) = S_F(\mathbf{q}, \mathbf{p}) = \frac{1}{2} (D_F(\mathbf{p}||\mathbf{q}) + D_F(\mathbf{q}||\mathbf{p})) = \frac{1}{2} \langle \mathbf{p} - \mathbf{q}, \mathbf{p}' - \mathbf{q}' \rangle.$$

An example of such a symmetrized divergence is the symmetric Kullback-Leibler divergence (SKL) widely used in computer vision and sound processing (see for example [29]).

A key observation is to note that the divergence S_F between two points of \mathcal{X} can be measured as a divergence in $\mathcal{X} \times \mathcal{X}' \subset \mathbb{R}^{2d}$. More precisely, let $\tilde{\mathbf{x}} = [\mathbf{x} \ \mathbf{x}']^T$ be the $2d$ -dimensional vector obtained by stacking the coordinates of \mathbf{x} on top of those of \mathbf{x}' , the gradient of F at \mathbf{x} . We have :

Theorem 1 $S_F(\mathbf{p}, \mathbf{q}) = \frac{1}{2} D_{\tilde{F}}(\tilde{\mathbf{p}}||\tilde{\mathbf{q}})$ where $\tilde{F}(\tilde{\mathbf{x}}) = F(\mathbf{x}) + F^*(\mathbf{x}')$ and $D_{\tilde{F}}$ is the Bregman divergence defined over $\mathcal{X} \times \mathcal{X}' \subset \mathbb{R}^{2d}$ for the generator function \tilde{F} .

Proof: Using Lemma 3, we have

$$S_F(\mathbf{p}, \mathbf{q}) = \frac{1}{2} (D_F(\mathbf{p}||\mathbf{q}) + D_F(\mathbf{q}||\mathbf{p})) = \frac{1}{2} (D_F(\mathbf{p}||\mathbf{q}) + D_{F^*}(\mathbf{p}'||\mathbf{q}')) = \frac{1}{2} D_{\tilde{F}}(\tilde{\mathbf{p}}||\tilde{\mathbf{q}})$$

□

It should be noted that $\tilde{\mathbf{x}}$ lies on the d -manifold $\tilde{\mathcal{X}} = \{[\mathbf{x} \ \mathbf{x}']^T \mid \mathbf{x} \in \mathbb{R}^d\}$ of \mathbb{R}^{2d} . Note also that $S_F(\mathbf{p}, \mathbf{q})$ is symmetric but *not* a Bregman divergence in general since $\tilde{\mathcal{X}}$ may not be convex, while $D_{\tilde{F}}$ is a non symmetric Bregman divergence in $\mathcal{X} \times \mathcal{X}'$.

2.4 Exponential families

2.4.1 Parametric statistical spaces and exponential families

A *statistical space* \mathcal{X} is an abstract space where coordinates of vector points $\boldsymbol{\theta} \in \mathcal{X}$ encode the parameters of statistical distributions. The dimension $d = \dim \mathcal{X}$ of the statistical space

coincides with the finite number of free parameters of the distribution laws. For example, the space $\mathcal{X} = \{[\mu \ \sigma]^T \mid (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+^+\}$ of univariate normal distributions $\mathcal{N}(\mu, \sigma)$ is a 2D parametric statistical space, extensively studied in information geometry [1] under the auspices of differential geometry. A prominent class of distribution families called the *exponential families* \mathcal{E}_F [1] admits the same *canonical* probability distribution function

$$p(x|\boldsymbol{\theta}) \stackrel{\text{def}}{=} \exp\{\langle \boldsymbol{\theta}, \mathbf{f}(x) \rangle - F(\boldsymbol{\theta}) + C(x)\}, \quad (7)$$

where $\mathbf{f}(x)$ denotes the *sufficient statistics* and $\boldsymbol{\theta} \in \mathcal{X}$ represents the *natural parameters*. Space \mathcal{X} is thus called the *natural parameter space* and, since $\log \int_x p(x|\boldsymbol{\theta}) dx = \log 1 = 0$, we have $F(\boldsymbol{\theta}) = \log \int_x \exp\{\langle \boldsymbol{\theta}, \mathbf{f}(\mathbf{x}) \rangle + C(x)\} dx$. F is called the *cumulant function* or the *log-partition function*. F fully characterizes the exponential family \mathcal{E}_F while term $C(x)$ ensures density normalization. (That is, $p(x|\boldsymbol{\theta})$ is indeed a probability density function satisfying $\int_x p(x|\boldsymbol{\theta}) dx = 1$.)

When the components of the sufficient statistics are affinely independent, this canonical representation is said to be *minimal*, and the family \mathcal{E}_F is called a *full* exponential family of order $d = \dim \mathcal{X}$. Moreover, we consider *regular* exponential families \mathcal{E}_F that have their support domains topologically open. Regular exponential families include many famous distribution laws such as Bernoulli (multinomial), Normal (univariate, multivariate and rectified), Poisson, Laplacian, negative binomial, Rayleigh, Wishart, Dirichlet, and Gamma distributions. Table 2 summarizes the various relevant parts of the canonical decompositions of some of these usual statistical distributions. Observe that the product of any two distributions of the same exponential family is another exponential family distribution that may not have anymore a nice parametric form (except for products of normal distribution pdfs that yield again normal distribution pdfs). Thus exponential families provide a unified treatment framework of common distributions. Note, however, that the uniform distribution *does not* belong to the exponential families.

2.4.2 Kullback-Leibler divergence of exponential families

In such statistical spaces \mathcal{X} , a basic primitive is to measure the *distortion* between any two distributions. The *Kullback-Leibler divergence* (also called *relative entropy* or information divergence, *I*-divergence) is a standard information-theoretic measure between two statistical distributions d_1 and d_2 defined as $\text{KL}(d_1||d_2) \stackrel{\text{def}}{=} \int_x d_1(x) \log \frac{d_1(x)}{d_2(x)} dx$. This statistical measure is not symmetric nor does the triangle inequality holds.

The link with Bregman divergences comes from the remarkable property that the Kullback-Leibler divergence of any two distributions of the *same* exponential family with respective natural parameters $\boldsymbol{\theta}_p$ and $\boldsymbol{\theta}_q$ is obtained from the Bregman divergence induced by the cumulant function of that family by *swapping* arguments. By a slight abuse of notations, we denote by $\text{KL}(\boldsymbol{\theta}_p||\boldsymbol{\theta}_q)$ the oriented Kullback-Leibler divergence between the probability density functions defined by the respective natural parameters, i.e. $\text{KL}(\boldsymbol{\theta}_p||\boldsymbol{\theta}_q) \stackrel{\text{def}}{=} \int_x p(x|\boldsymbol{\theta}_p) \log \frac{p(x|\boldsymbol{\theta}_p)}{p(x|\boldsymbol{\theta}_q)} dx$.

Exponential family			
Canonical probability density function: $\exp\{\langle \boldsymbol{\theta}, \mathbf{f}(x) \rangle - F(\boldsymbol{\theta}) + C(x)\}$			
Natural parameters $\boldsymbol{\theta}$	Sufficient statistics $\mathbf{f}(x)$	Cumulant function $F(\boldsymbol{\theta})$	Dens. Norm. $C(x)$
Bernoulli $\mathcal{B}(q)$ (Tossing coin with $\Pr(\text{heads}) = q$ and $\Pr(\text{tails}) = 1 - q$)			
$\log \frac{q}{1-q}$	x	$\log(1 + \exp \theta)$	0
Multinomial $\mathcal{M}(q_1, \dots, q_{d+1})$ (Extend Bernoulli with $\Pr(x_i) = q_i$ and $\sum_i q_i = 1$)			
$\theta_i = \log \frac{q_i}{1 - \sum_{j=1}^d q_j}$	$f_i(\mathbf{x}) = x_i$	$\log(1 + \sum_{i=1}^d \exp \theta_i)$	0
Beta $\beta(\theta_1, \theta_2)$ (Bernoulli conjugate prior)			
$[\theta_1 \ \theta_2]^T$	$[\log x \ \log(1-x)]^T$	$\log B(\theta_1 + 1, \theta_2 + 1)$	0
$F(\boldsymbol{\theta}) = \log \frac{\Gamma(\theta_1+1)\Gamma(\theta_2+1)}{\Gamma(\theta_1+\theta_2+2)}$ (with $\Gamma(x) = \int_0^\infty t^{x-1} \exp(-t) dt = (x-1)\Gamma(x-1)$)			
Univariate Normal $\mathcal{N}(\mu, \sigma^2)$			
$[\frac{\mu}{\sigma^2} \ \frac{-1}{2\sigma^2}]^T$	$[x \ x^2]^T$	$-\frac{\theta_1^2}{4\theta_2} + \frac{1}{2} \log(-\frac{\pi}{\theta_2})$	0
Multivariate Normal $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$			
$[\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \ -\frac{1}{2}\boldsymbol{\Sigma}^{-1}]$	$[\mathbf{x} \ \mathbf{x}\mathbf{x}^T]$	$\frac{1}{2}\boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \frac{1}{2} \log \det(2\pi\boldsymbol{\Sigma})$	0
Rayleigh $\mathcal{R}(\sigma^2)$ (used in ultrasound imageries)			
$-\frac{1}{2\sigma^2}$	x^2	$\log -\frac{1}{2\theta}$	$\log x$
Laplacian $\mathcal{L}(\theta)$ (used in radioactivity decay)			
θ	$-x$	$-\log \theta$	0
Poisson $\mathcal{P}(\lambda)$ (counting process)			
$\log \lambda$	x	$\exp \theta$	$-\log x!$
Gamma $\gamma(\theta_1, \theta_2)$ (waiting times in Poisson processes)			
$[\theta_1 \ \theta_2]^T$	$[\log x \ x]^T$	$\log \Gamma(\theta_1 + 1) + (\theta_2 + 1) \log(-\theta_2)$	0
Dirichlet $\mathcal{D}(\boldsymbol{\alpha})$ (varying proportion model $\ \mathbf{x}\ = 1$, conjugate prior of Multinomial)			
$\theta_i = \alpha_i - 1$	$f_i(\mathbf{x}) = \log x_i$	$\log \Gamma(\sum_i \theta_i + d) - \sum_i \Gamma(\theta_i + 1)$	0

Table 2: Canonical decompositions of usual exponential families.

The following theorem is the extension to the continuous case of a result mentioned in [6].

Theorem 2 *The Kullback-Leibler divergence of any two distributions of the same exponential family with natural parameters $\boldsymbol{\theta}_p$ and $\boldsymbol{\theta}_q$ is obtained from the Bregman divergence induced by the cumulant function F as: $\text{KL}(\boldsymbol{\theta}_p||\boldsymbol{\theta}_q) = D_F(\boldsymbol{\theta}_q||\boldsymbol{\theta}_p)$.*

Before proving the theorem, we note that

$$\nabla F(\boldsymbol{\theta}) = \left[\int_x \mathbf{f}(x) \exp\{\langle \boldsymbol{\theta}, \mathbf{f}(x) \rangle - F(\boldsymbol{\theta}) + C(x)\} dx \right]. \quad (8)$$

The coordinates of $\boldsymbol{\mu} \stackrel{\text{def}}{=} \nabla F(\boldsymbol{\theta}) = [\int_x \mathbf{f}(\mathbf{x}) p(x|\boldsymbol{\theta}) dx] = E_{\boldsymbol{\theta}}(\mathbf{f}(\mathbf{x}))$ are called the *expectation parameters*. As an example, consider the univariate normal distribution $\mathcal{N}(\mu, \sigma)$ with sufficient statistics $[x \ x^2]^T$ (see Table 2). The expectation parameters are $\boldsymbol{\mu} = \nabla F(\boldsymbol{\theta}) = [\mu \ \mu^2 + \sigma^2]^T$, where $\mu = \int_x x p(x|\boldsymbol{\theta}) dx$ and $\mu^2 + \sigma^2 = \int_x x^2 p(x|\boldsymbol{\theta}) dx$.

We now prove the theorem.

Proof:

$$\begin{aligned} \text{KL}(\boldsymbol{\theta}_p||\boldsymbol{\theta}_q) &= \int_x p(x|\boldsymbol{\theta}_p) \log \frac{p(x|\boldsymbol{\theta}_p)}{p(x|\boldsymbol{\theta}_q)} dx \\ &= \int_x p(x|\boldsymbol{\theta}_p) (F(\boldsymbol{\theta}_q) - F(\boldsymbol{\theta}_p) + \langle \boldsymbol{\theta}_p - \boldsymbol{\theta}_q, \mathbf{f}(x) \rangle) dx \\ &= \int_x p(x|\boldsymbol{\theta}_p) (D_F(\boldsymbol{\theta}_q||\boldsymbol{\theta}_p) + \langle \boldsymbol{\theta}_q - \boldsymbol{\theta}_p, \nabla F(\boldsymbol{\theta}_p) \rangle + \langle \boldsymbol{\theta}_p - \boldsymbol{\theta}_q, \mathbf{f}(x) \rangle) dx \\ &= D_F(\boldsymbol{\theta}_q||\boldsymbol{\theta}_p) + \int_x p(x|\boldsymbol{\theta}_p) \langle \boldsymbol{\theta}_q - \boldsymbol{\theta}_p, \nabla F(\boldsymbol{\theta}_p) - \mathbf{f}(x) \rangle dx \\ &= D_F(\boldsymbol{\theta}_q||\boldsymbol{\theta}_p) - \int_x p(x|\boldsymbol{\theta}_p) \langle \boldsymbol{\theta}_q - \boldsymbol{\theta}_p, \mathbf{f}(x) \rangle dx + \langle \boldsymbol{\theta}_q - \boldsymbol{\theta}_p, \nabla F(\boldsymbol{\theta}_p) \rangle \\ &\stackrel{(\text{Eq. 8})}{=} D_F(\boldsymbol{\theta}_q||\boldsymbol{\theta}_p) \end{aligned}$$

□

2.4.3 Dual parameterizations and dual divergences

The notion of dual Bregman divergences introduced earlier and dual parameterizations extend naturally to statistical spaces. Since, $\boldsymbol{\mu} = \nabla F(\boldsymbol{\theta})$ (Eq. 8), the convex conjugate of $F(\boldsymbol{\theta})$ is $F^*(\boldsymbol{\mu}) = \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle - F(\boldsymbol{\theta})$ (Eq. 6). From Lemma 3, we then deduce the following theorem.

Theorem 3 $D_F(\boldsymbol{\theta}_p||\boldsymbol{\theta}_q) = D_{F^*}(\boldsymbol{\mu}_q||\boldsymbol{\mu}_p)$ where F^* denote the convex conjugate of F .

Bernoulli dual divergences: Logistic loss/binary relative entropy		
$F(\theta) = \log(1 + \exp \theta)$	$D_F(\theta \theta') = \log \frac{1+\exp \theta}{1+\exp \theta'} - (\theta - \theta') \frac{\exp \theta'}{1+\exp \theta'}$	$f(\theta) = \frac{\exp \theta}{1+\exp \theta} = \mu$
$F^*(\mu) = \mu \log \mu + (1 - \mu) \log(1 - \mu)$	$D_{F^*}(\mu' \mu) = \mu' \log \frac{\mu'}{\mu} + (1 - \mu) \log \frac{1-\mu'}{1-\mu}$	$f^*(\mu) = \log \frac{\mu}{1-\mu} = \theta$
Poisson dual divergences: Exponential loss/Unnormalized Shannon entropy		
$F(\theta) = \exp \theta$	$D_F(\theta \theta') = \exp \theta - \exp \theta' - (\theta - \theta') \exp \theta'$	$f(\theta) = \exp \theta = \mu$
$F^*(\mu) = \mu \log \mu - \mu$	$D_{F^*}(\mu' \mu) = \mu' \log \frac{\mu'}{\mu} + \mu - \mu'$	$f^*(\mu) = \log \mu = \theta$

Table 3: Examples of dual parameterizations of exponential families and their corresponding Kullback-Leibler (Bregman) divergences for the Bernoulli and Poisson distributions.

Table 3 presents some examples of dual parameterizations of exponential families (i.e., the natural $\boldsymbol{\theta}$ -parameters and expectation $\boldsymbol{\mu}$ -parameters and dual Legendre cumulant functions), and describe the corresponding Bregman divergences induced by the Kullback-Leibler divergences.

Finally, we would like to point out that Banerjee et al. [6] have shown that there is a *bijection* between the regular exponential families and a subset of the Bregman divergences called *regular Bregman divergences*.

3 Elements of Bregman geometry

In this section, we discuss several basic geometric properties that will be useful when studying Bregman Voronoi diagrams. Specifically, we characterize Bregman bisectors, Bregman balls and Bregman geodesics. Since Bregman divergences are not symmetric, we describe several types of Bregman bisectors in §3.1. We subsequently characterize Bregman balls by using a lifting transform that extends a construction well-known in the Euclidean case (§3.2). Finally, we characterize geodesics and show an orthogonality property between bisectors and geodesics in §3.3.

3.1 Bregman bisectors

Since Bregman divergences are not symmetric, we can define several types of bisectors. The Bregman bisector of the *first type* is defined as

$$H_F(\mathbf{p}, \mathbf{q}) = \{\mathbf{x} \in \mathcal{X} \mid D_F(\mathbf{x}||\mathbf{p}) = D_F(\mathbf{x}||\mathbf{q})\}.$$

Similarly, we define the Bregman bisector of the *second type* as

$$H'_F(\mathbf{p}, \mathbf{q}) = \{\mathbf{x} \in \mathcal{X} \mid D_F(\mathbf{p}||\mathbf{x}) = D_F(\mathbf{q}||\mathbf{x})\}.$$

These bisectors are identical when the divergence is symmetric. However, in general, they are distinct, the bisectors of the first type being linear while the bisectors of the second type are potentially curved (but always linear in the gradient space, hence the notation). More precisely, we have the following lemma

Lemma 4 *The Bregman bisector of the first type $H_F(\mathbf{p}, \mathbf{q})$ is the hyperplane of equation:*

$$H_F(\mathbf{p}, \mathbf{q}) : \langle \mathbf{x}, \mathbf{p}' - \mathbf{q}' \rangle + F(\mathbf{p}) - \langle \mathbf{p}, \mathbf{p}' \rangle - F(\mathbf{q}) + \langle \mathbf{q}, \mathbf{q}' \rangle = 0$$

The Bregman bisector of the second type $H'_F(\mathbf{p}, \mathbf{q})$ is the hypersurface of equation

$$H'_F(\mathbf{p}, \mathbf{q}) : \langle \mathbf{x}', \mathbf{q} - \mathbf{p} \rangle + F(\mathbf{p}) - F(\mathbf{q}) = 0$$

(a hyperplane in the gradient space \mathcal{X}').

It should be noted that \mathbf{p} and \mathbf{q} lie necessarily on different sides of $H_F(\mathbf{p}, \mathbf{q})$ since $H_F(\mathbf{p}, \mathbf{q})(\mathbf{p}) = -D_F(\mathbf{p}||\mathbf{q}) < 0$ and $H_F(\mathbf{p}, \mathbf{q})(\mathbf{q}) = D_F(\mathbf{q}||\mathbf{p}) > 0$.

From Lemma 3, we know that $D_F(\mathbf{x}||\mathbf{y}) = D_{F^*}(\mathbf{y}'||\mathbf{x}')$ where F^* is the convex conjugate of F . We therefore have

$$\begin{aligned} H_F(\mathbf{p}, \mathbf{q}) &= \nabla^{-1} F(H'_{F^*}(\mathbf{q}', \mathbf{p}')), \\ H'_F(\mathbf{p}, \mathbf{q}) &= \nabla^{-1} F(H_{F^*}(\mathbf{q}', \mathbf{p}')). \end{aligned}$$

Figure 4 depicts several first-type and second-type bisectors for various pairs of primal/dual Bregman divergences.

The bisector $H''_F(\mathbf{p}, \mathbf{q})$ for the *symmetrized Bregman divergence* S_F is given by

$$H''_F(\mathbf{p}, \mathbf{q}) : \langle \mathbf{x}, \mathbf{q}' - \mathbf{p}' \rangle + \langle \mathbf{x}', \mathbf{q} - \mathbf{p} \rangle + \langle \mathbf{p}, \mathbf{p}' \rangle - \langle \mathbf{q}, \mathbf{q}' \rangle = 0.$$

Such a bisector is not linear in \mathbf{x} nor in \mathbf{x}' . However, we can observe that the expression is linear in $\tilde{\mathbf{x}} = [\mathbf{x} \ \mathbf{x}']^T$. Indeed, proceeding as we did in §2.3, we can rewrite the above equation as

$$H_{\tilde{F}}(\tilde{\mathbf{p}}, \tilde{\mathbf{q}}) : \left\langle \begin{bmatrix} \mathbf{x} \\ \mathbf{x}' \end{bmatrix}, \begin{bmatrix} \mathbf{q}' - \mathbf{p}' \\ \mathbf{q} - \mathbf{p} \end{bmatrix} \right\rangle + \langle \mathbf{p}, \mathbf{p}' \rangle - \langle \mathbf{q}, \mathbf{q}' \rangle = 0.$$

which shows that $H''_F(\mathbf{p}, \mathbf{q})$ is the projection on \mathcal{X} of the intersection of the hyperplane $H(\tilde{\mathbf{p}}, \tilde{\mathbf{q}})$ of \mathbb{R}^{2d} with the d -dimensional manifold $\tilde{\mathcal{X}} = \{\tilde{\mathbf{x}} = [\mathbf{x} \ \mathbf{x}']^T \mid \mathbf{x} \in \mathcal{X}\}$.

3.2 Bregman spheres and the lifting map

We define the Bregman balls of, respectively, the first and the second types as

$$B_F(\mathbf{c}, r) = \{\mathbf{x} \in \mathcal{X} \mid D_F(\mathbf{x}||\mathbf{c}) \leq r\} \quad \text{and} \quad B'_F(\mathbf{c}, r) = \{\mathbf{x} \in \mathcal{X} \mid D_F(\mathbf{c}||\mathbf{x}) \leq r\}$$

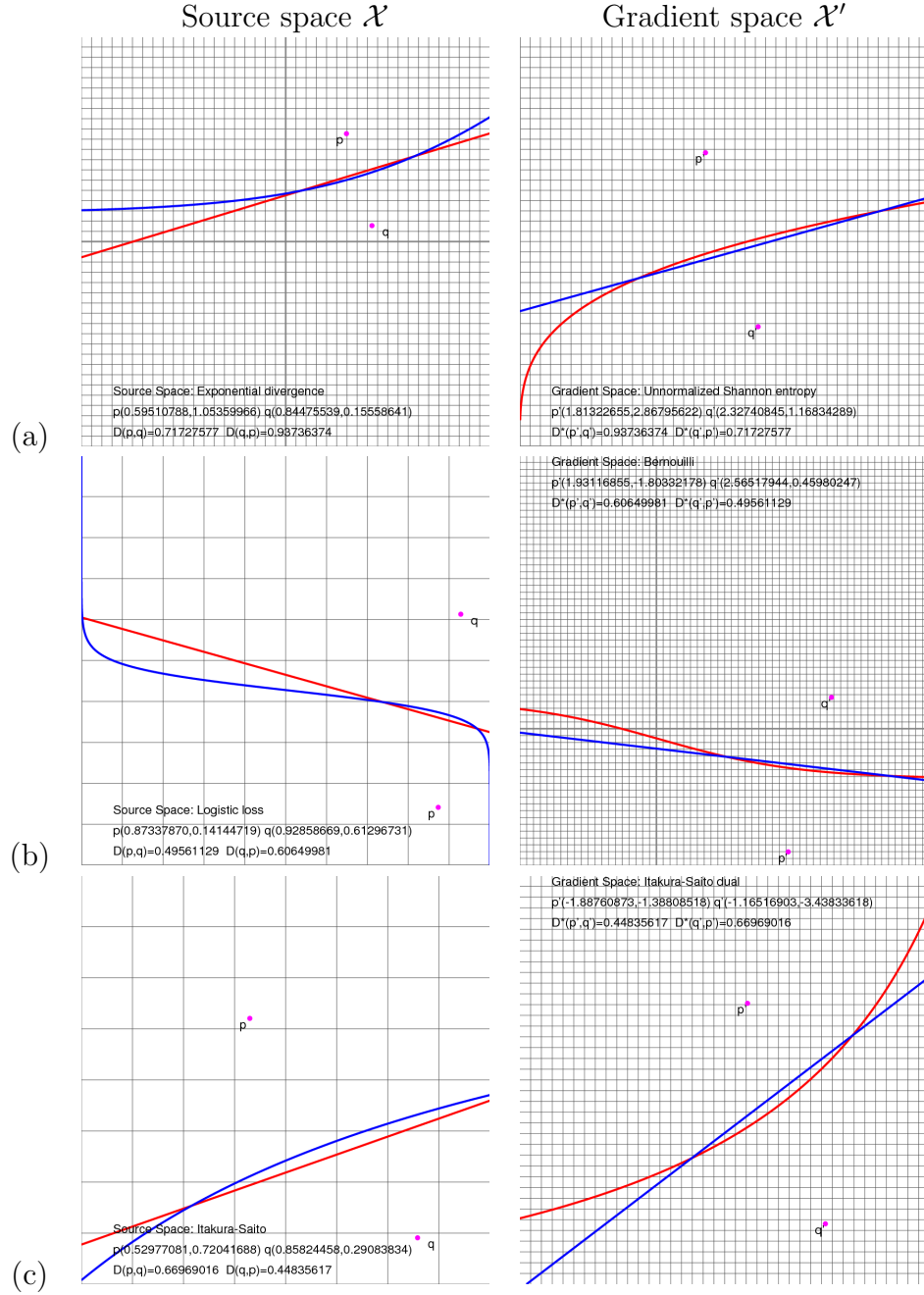


Figure 4: Bregman bisectors: first-type linear bisector and second-type curved bisector are displayed for pairs of primal/dual Bregman divergences: (a) exponential loss/unnormalized Shannon entropy, (b) logistic loss/dual logistic loss, and (c) self-dual Itakura-Saito divergence. The grid size of \mathbb{R}^2 in \mathcal{X} and \mathcal{X}' is ten ticks per unit. First-type (primal linear/dual curved) and second-type (primal curved/dual linear) bisectors are respectively drawn in red and blue.

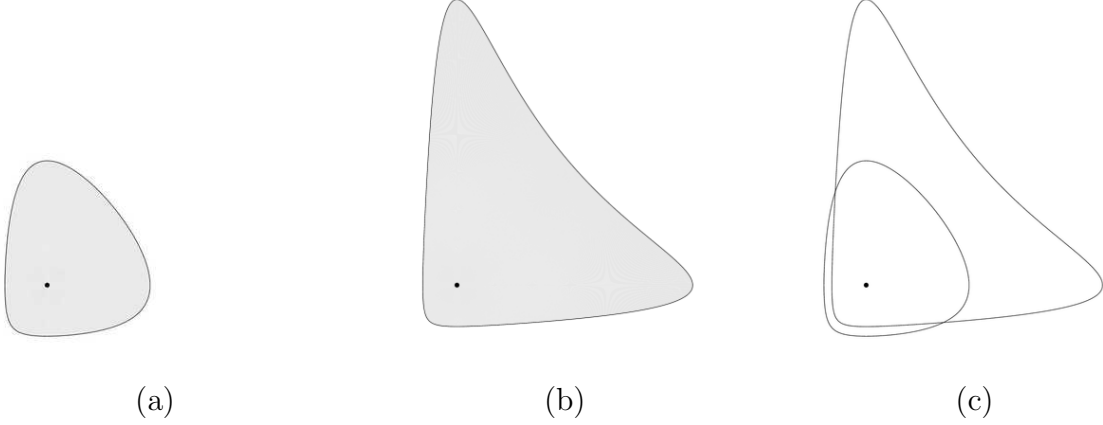


Figure 5: Bregman balls for the Itakura-Saito divergence. The (convex) ball (a) of the first type $B_F(\mathbf{c}, r)$, (b) the ball of the second type $B'_F(\mathbf{c}, r)$ with the same center and radius, (c) superposition of the two corresponding bounding spheres.

The Bregman balls of the first type are convex while this is not necessarily true for the balls of the second type as shown in Fig. 5 for the Itakura-Saito divergence (defined in Table 1). The associated bounding *Bregman spheres* are obtained by replacing the inequalities by equalities.

From Lemma 3, we deduce that

$$B'_F(\mathbf{c}, r) = \nabla^{-1} F(B_{F^*}(\mathbf{c}', r)). \quad (9)$$

Let us now examine a few properties of Bregman spheres using a lifting transformation that generalizes a similar construct for Euclidean spheres (see [10, 33]).

Let us embed the domain \mathcal{X} in $\hat{\mathcal{X}} = \mathcal{X} \times \mathbb{R} \subset \mathbb{R}^{d+1}$ using an *extra dimension* denoted by the Z -axis. For a point $\mathbf{x} \in \mathcal{X}$, recall that $\hat{\mathbf{x}} = (\mathbf{x}, F(\mathbf{x}))$ denotes the point obtained by lifting \mathbf{x} onto \mathcal{F} (see Figure 1). In addition, write $\text{Proj}_{\mathcal{X}}(\mathbf{x}, z) = \mathbf{x}$ for the projection of a point of $\hat{\mathcal{X}}$ onto \mathcal{X} .

Let $\mathbf{p} \in \mathcal{X}$ and $H_{\mathbf{p}}$ be the hyperplane tangent to \mathcal{F} at point $\hat{\mathbf{p}}$ of equation

$$z = H_{\mathbf{p}}(\mathbf{x}) = \langle \mathbf{x} - \mathbf{p}, \mathbf{p}' \rangle + F(\mathbf{p}),$$

and let $H_{\mathbf{p}}^{\uparrow}$ denote the halfspace above $H_{\mathbf{p}}$ consisting of the points $\mathbf{x} = [\mathbf{x} \ z]^T \in \hat{\mathcal{X}}$ such that $z > H_{\mathbf{p}}(\mathbf{x})$. Let $\sigma(\mathbf{c}, r)$ denote either the first-type or second-type Bregman sphere centered at \mathbf{c} with radius r (i.e., $\partial B_F(\mathbf{c}, r)$ or $\partial B'_F(\mathbf{c}, r)$).

The lifted image $\hat{\sigma}$ of a Bregman sphere σ is $\hat{\sigma} = \{(\mathbf{x}, F(\mathbf{x})), \mathbf{x} \in \sigma\}$. We associate to a Bregman sphere $\sigma = \sigma(\mathbf{c}, r)$ of \mathcal{X} the hyperplane

$$H_{\sigma} : z = \langle \mathbf{x} - \mathbf{c}, \mathbf{c}' \rangle + F(\mathbf{c}) + r, \quad (10)$$

parallel to $H_{\mathbf{c}}$ and at vertical distance r from $H_{\mathbf{c}}$ (see Figure 6). Observe that H_{σ} coincides with $H_{\mathbf{c}}$ when $r = 0$, i.e. when sphere σ is reduced to a single point.

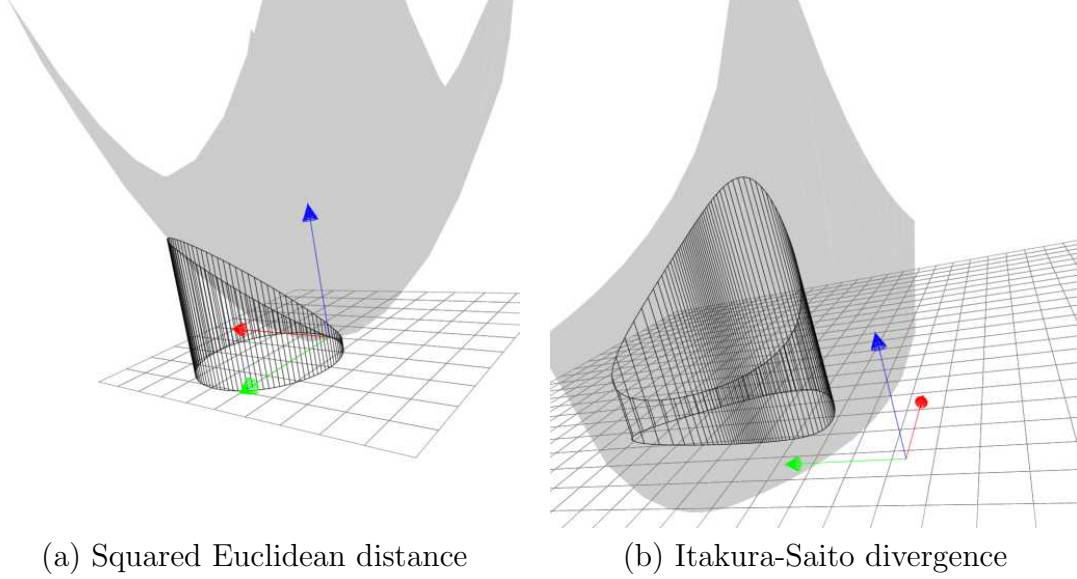


Figure 6: Two Bregman circles σ and the associated curves $\hat{\sigma}$ obtained by lifting σ onto \mathcal{F} . The curves $\hat{\sigma}$ are obtained as the intersection of the hyperplane H_σ with the convex hypersurface \mathcal{F} . 3D illustration with (a) the squared Euclidean distance, and (b) the Itakura-Saito divergence.

Lemma 5 *$\hat{\sigma}$ is the intersection of \mathcal{F} with H_σ . Conversely, the intersection of any hyperplane H with \mathcal{F} projects onto \mathcal{X} as a Bregman sphere. More precisely, if the equation of H is $z = \langle \mathbf{x}, \mathbf{a} \rangle + b$, the sphere is centered at $\mathbf{c} = \nabla^{-1}F(\mathbf{a})$ and its radius is $\langle \mathbf{a}, \mathbf{c} \rangle - F(\mathbf{c}) + b$.*

Proof: The first part of the lemma is a direct consequence of the fact that $D_F(\mathbf{x}||\mathbf{y})$ is measured by the vertical distance from $\hat{\mathbf{x}}$ to $H_{\mathbf{y}}$ (see Lemma 1). For the second part, we consider the hyperplane H^\parallel parallel to H and tangent to \mathcal{F} . From Eq. 10, we deduce $\mathbf{a} = \mathbf{c}'$. The equation of H^\parallel is thus $z = \langle \mathbf{x} - \nabla^{-1}F(\mathbf{a}), \mathbf{a} \rangle + F(\nabla^{-1}F(\mathbf{a}))$. It follows that the divergence from any point of σ to \mathbf{c} , which is equal to the vertical distance between H and H^\parallel , is $\langle \nabla^{-1}F(\mathbf{a}), \mathbf{a} \rangle - F(\nabla^{-1}F(\mathbf{a})) + b = \langle \mathbf{a}, \mathbf{c} \rangle - F(\mathbf{c}) + b$. \square

Bregman spheres have been defined as manifolds of codimension 1 of \mathbb{R}^d , i.e. hyperspheres. More generally, we can define the Bregman spheres of codimension $k+1$ of \mathbb{R}^d as the Bregman (hyper)spheres of some affine space $\mathcal{Z} \subset \mathbb{R}^d$ of codimension k . The next lemma shows that Bregman spheres are stable under intersection.

Lemma 6 *The intersection of k Bregman spheres $\sigma_1, \dots, \sigma_k$ is a Bregman sphere σ . If the σ_i pairwise intersect transversally, $\sigma = \cap_{i=1}^k \sigma_i$ is a k -Bregman sphere.*

Proof: Consider first the case of Bregman spheres of the first type. The k hyperplanes H_{σ_i} , $i = 1, \dots, k$ intersect along an affine space H of codimension k of \mathbb{R}^{d+1} that vertically projects onto G . Let $G^\uparrow = G \times \mathbb{R}$ be the vertical flat of codimension k that contains G (and

H) and write $\mathcal{F}_G = \mathcal{F} \cap G^\uparrow$ and $H_G = H \cap G^\uparrow$. Note that \mathcal{F}_G is the graph of the restriction of F to G and that H_G is a hyperplane of G^\uparrow . We can therefore apply Lemma 5 in G^\uparrow , which proves the lemma for Bregman spheres of the first type.

The case of Bregman spheres of the second type follows from the duality of Eq. 9. \square

Union and intersection of Bregman balls

Theorem 4 *The union of n Bregman balls has combinatorial complexity $\Theta(n^{\lfloor \frac{d+1}{2} \rfloor})$ and can be computed in time $\Theta(n \log n + n^{\lfloor \frac{d+1}{2} \rfloor})$.*

Proof: To each ball, we can associate its bounding Bregman sphere σ_i which, by Lemma 5, is the projection by $\text{Proj}_{\mathcal{X}}$ of the intersection of \mathcal{F} with a hyperplane H_{σ_i} . The points of \mathcal{F} that are below H_{σ_i} projects onto points that are inside the Bregman ball bounded by σ_i . Hence, the union of the balls is the projection by $\text{Proj}_{\mathcal{X}}$ of the complement of $\mathcal{F} \cap \mathcal{H}^\uparrow$ where $\mathcal{H}^\uparrow = \cap_{i=1}^n H_{\sigma_i}^\uparrow$. \mathcal{H}^\uparrow is a convex polytope defined as the intersection of n half-spaces. The theorem follows from McMullen's theorem that bounds the number of faces of a polytope [31], and Chazelle's optimal convex hull/half-space intersection algorithm [14]. The result for the balls of the second type is deduced from the result for the balls of the first type and the duality of Eq. 9. \square

Very similar arguments prove the following theorem (just replace $H_{\sigma_i}^\uparrow$ by the complementary halfspace $H_{\sigma_i}^\downarrow$).

Theorem 5 *The intersection of n Bregman balls has combinatorial complexity $\Theta(n^{\lfloor \frac{d+1}{2} \rfloor})$ and can be computed in time $\Theta(n \log n + n^{\lfloor \frac{d+1}{2} \rfloor})$.*

Circumscribing Bregman spheres. There exists, in general, a unique Bregman sphere passing through $d + 1$ points of \mathbb{R}^d . This is easily shown using the lifting map since, in general, there exists a unique hyperplanes of \mathbb{R}^{d+1} passing through $d + 1$ points. The claim then follows from Lemma 5.

Deciding whether a point \mathbf{x} falls *inside*, *on* or *outside* a Bregman sphere $\sigma \in \mathbb{R}^d$ passing through $d + 1$ points of $\mathbf{p}_0, \dots, \mathbf{p}_d$ will be crucial for computing Bregman Voronoi diagrams and associated triangulations. The lifting map immediately implies that such a decision task reduces to determining the orientation of the simplex $(\hat{\mathbf{p}}_0, \dots, \hat{\mathbf{p}}_d, \hat{\mathbf{x}})$ of \mathbb{R}^{d+1} , which in turn reduces to evaluating the sign of the determinant of the $(d + 2) \times (d + 2)$ matrix (see [32])

$$\text{InSphere}(\mathbf{x}; \mathbf{p}_0, \dots, \mathbf{p}_d) = \begin{vmatrix} 1 & \dots & 1 & 1 \\ \mathbf{p}_0 & \dots & \mathbf{p}_d & \mathbf{x} \\ F(\mathbf{p}_0) & \dots & F(\mathbf{p}_d) & F(\mathbf{x}) \end{vmatrix}$$

If one assumes that the determinant $\begin{vmatrix} 1 & \dots & 1 \\ \mathbf{p}_0 & \dots & \mathbf{p}_d \end{vmatrix}$ is non-zero, $\text{InSphere}(\mathbf{x}; \mathbf{p}_0, \dots, \mathbf{p}_d)$ is negative, null or positive depending on whether \mathbf{x} lies inside, on, or outside σ .

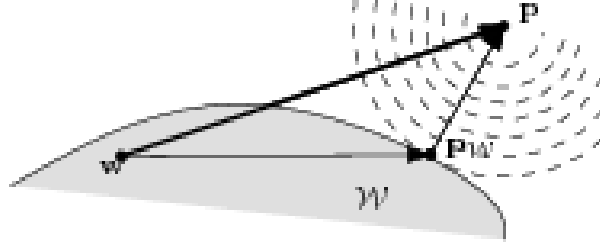


Figure 7: Generalized Pythagoras' theorem for Bregman divergences: The projection \mathbf{p}_W of point \mathbf{p} to a convex subset $W \subseteq \mathcal{X}$. For convex subset W , we have $D_F(\mathbf{w}||\mathbf{p}) \geq D_F(\mathbf{w}||\mathbf{p}_W) + D_F(\mathbf{p}_W||\mathbf{p})$ (with equality for and only for affine sets W).

3.3 Projection, orthogonality and geodesics

We start with an easy property of Bregman divergences.

Property 5 (Three-point property) *For any triple \mathbf{p}, \mathbf{q} and \mathbf{r} of points of \mathcal{X} , we have:*
 $D_F(\mathbf{p}||\mathbf{q}) + D_F(\mathbf{q}||\mathbf{r}) = D_F(\mathbf{p}||\mathbf{r}) + \langle \mathbf{p} - \mathbf{q}, \mathbf{r}' - \mathbf{q}' \rangle$.

The following lemma characterizes the *Bregman projection* of a point onto a closed convex set W .

Lemma 7 (Bregman projection) *For any \mathbf{p} , there exists a unique point $\mathbf{x} \in W$ that minimizes $D_F(\mathbf{x}||\mathbf{p})$. We call this point the Bregman projection of \mathbf{p} onto W and denote it \mathbf{p}_W .*

Proof: If it is not the case, then define \mathbf{x} and \mathbf{y} two minimizers with $D_F(\mathbf{x}||\mathbf{p}) = D_F(\mathbf{y}||\mathbf{p}) = l$. Since W is convex, $(\mathbf{x} + \mathbf{y})/2 \in W$ and, since D_F is strictly convex in its first argument (see Section 2.1), $D_F((\mathbf{x} + \mathbf{y})/2||\mathbf{p}) < D_F(\mathbf{x}||\mathbf{p})/2 + D_F(\mathbf{y}||\mathbf{p})/2$. But $D_F(\mathbf{x}||\mathbf{p})/2 + D_F(\mathbf{y}||\mathbf{p})/2 = l$ yielding a contradiction. \square

We now introduce the notion of *Bregman orthogonality*. We say that \mathbf{pq} is Bregman orthogonal to \mathbf{qr} iff $D_F(\mathbf{p}||\mathbf{q}) + D_F(\mathbf{q}||\mathbf{r}) = D_F(\mathbf{p}||\mathbf{r})$ or equivalently (by the Three-point property), if and only if $\langle \mathbf{p} - \mathbf{q}, \mathbf{r}' - \mathbf{q}' \rangle = 0$. Observe the analogy with Pythagoras' theorem in Euclidean space (see Figure 7). Note also that the orthogonality relation is not symmetric: the fact that \mathbf{pq} is Bregman orthogonal to \mathbf{qr} does not necessarily imply that \mathbf{qr} is Bregman orthogonal to \mathbf{pq} . More generally, we say that $I \subseteq \mathcal{X}$ is *Bregman orthogonal* to $J \subseteq \mathcal{X}$ ($I \cap J \neq \emptyset$) iff for any $\mathbf{p} \in I$ and $\mathbf{r} \in J$, there exists a $\mathbf{q} \in I \cap J$ such that \mathbf{pq} is Bregman orthogonal to \mathbf{qr} .

Notice that orthogonality is preserved in the gradient space. Indeed, since $\langle \mathbf{p} - \mathbf{q}, \mathbf{r}' - \mathbf{q}' \rangle = \langle \mathbf{r}' - \mathbf{q}', \mathbf{p} - \mathbf{q} \rangle$, \mathbf{pq} is Bregman orthogonal to \mathbf{qr} iff $\mathbf{r}'\mathbf{q}'$ is Bregman orthogonal to $\mathbf{q}'\mathbf{p}'$.

Let $\Gamma_F(\mathbf{p}, \mathbf{q})$ be the image by $\nabla^{-1}F$ of the line segment $\mathbf{p}'\mathbf{q}'$, i.e.

$$\Gamma_F(\mathbf{p}, \mathbf{q}) = \{\mathbf{x} \in \mathcal{X} : \mathbf{x}' = (1 - \lambda)\mathbf{p}' + \lambda\mathbf{q}', \lambda \in [0, 1]\}.$$

By analogy, we rename the line segment \mathbf{pq} as

$$\Lambda(\mathbf{p}, \mathbf{q}) = \{\mathbf{x} \in \mathcal{X} : \mathbf{x} = (1 - \lambda)\mathbf{p} + \lambda\mathbf{q}, \lambda \in [0, 1]\}$$

In the Euclidean case ($F(x) = \frac{1}{2}\|\mathbf{x}\|^2$), $\Gamma_F(\mathbf{p}, \mathbf{q}) = \Lambda(\mathbf{p}, \mathbf{q})$ is the unique geodesic path joining \mathbf{p} to \mathbf{q} and it is orthogonal to the bisector $H_F(\mathbf{p}, \mathbf{q})$. For general Bregman divergences, we have similar properties as shown next.

Lemma 8 $\Gamma_F(\mathbf{p}, \mathbf{q})$ is Bregman orthogonal to the Bregman bisector $H_F(\mathbf{p}, \mathbf{q})$ while $\Lambda(\mathbf{p}, \mathbf{q})$ is Bregman orthogonal to $H_{F^*}(\mathbf{p}, \mathbf{q})$.

Proof: Since \mathbf{p} and \mathbf{q} lie on different sides of $H_F(\mathbf{p}, \mathbf{q})$, $\Gamma_F(\mathbf{p}, \mathbf{q})$ must intersect $H_F(\mathbf{p}, \mathbf{q})$. Fix any distinct $\mathbf{x} \in \Gamma(\mathbf{p}, \mathbf{q})$ and $\mathbf{y} \in H_F(\mathbf{p}, \mathbf{q})$, and let $\mathbf{t} \in \Gamma(\mathbf{p}, \mathbf{q}) \cap H_F(\mathbf{p}, \mathbf{q})$. To prove the first part of the lemma, we need to show that $\langle \mathbf{y} - \mathbf{t}, \mathbf{x}' - \mathbf{t}' \rangle = 0$.

Since \mathbf{t} and \mathbf{x} both belong to $\Gamma_F(\mathbf{p}, \mathbf{q})$, we have $\mathbf{t}' - \mathbf{x}' = \lambda(\mathbf{p}' - \mathbf{q}')$, for some $\lambda \in \mathbb{R}$, and, since \mathbf{y} and \mathbf{t} belong to $H_F(\mathbf{p}, \mathbf{q})$, we deduce from the equation of $H_F(\mathbf{p}, \mathbf{q})$ that $\langle \mathbf{y} - \mathbf{t}, \mathbf{p}' - \mathbf{q}' \rangle = 0$. We conclude that $\langle \mathbf{y} - \mathbf{t}, \mathbf{x}' - \mathbf{t}' \rangle = 0$, which proves that $\Gamma_F(\mathbf{p}, \mathbf{q})$ is indeed Bregman orthogonal to $H_F(\mathbf{p}, \mathbf{q})$.

The second part of the lemma is easily proved by using the fact that orthogonality is preserved in the gradient space as noted above. \square

Figure 8 shows Bregman bisectors and their relationships with respect to $\Lambda(\mathbf{p}, \mathbf{q})$ and $\Gamma_F(\mathbf{p}, \mathbf{q})$.

We now focus on characterizing Bregman geodesics. First, recall that a parameterized curve \mathcal{C} between two points \mathbf{p}_0 and \mathbf{p}_1 is defined as a set $\mathcal{C} = \{\mathbf{p}_\lambda\}_{\lambda=0}^1$, which is continuous. In Riemannian geometry, geodesics are the curves that minimize the arc length with respect to the Riemannian metric [1, 27]. Since embedding \mathcal{X} with a Bregman divergence does not yield a metric space, we define the following curve lengths:

$$\ell_\Gamma(\mathcal{C}) = \int_{\lambda=0}^1 D_F(\mathbf{p}_0 || \mathbf{p}_\lambda) d\lambda, \quad (11)$$

$$\ell_\Lambda(\mathcal{C}) = \int_{\lambda=0}^1 D_F(\mathbf{p}_\lambda || \mathbf{p}_0) d\lambda. \quad (12)$$

We now characterize the dual pair of geodesics and their lengths as follows:

Lemma 9 Curve $\Gamma_F(\mathbf{p}_0, \mathbf{p}_1)$ (respectively straight line segment $\Lambda(\mathbf{p}_0, \mathbf{p}_1)$) minimizes $\int_{\lambda=0}^1 D_F(\mathbf{p}_0 || \mathbf{p}_\lambda) d\lambda$ (respectively $\int_{\lambda=0}^1 D_F(\mathbf{p}_\lambda || \mathbf{p}_0) d\lambda$) over all curves $\mathcal{C} = \{\mathbf{p}_\lambda\}_{\lambda=0}^1$.

Proof: For any curve \mathcal{C} between \mathbf{p}_0 and \mathbf{p}_1 , we measure the ℓ_Γ length as $\ell_\Gamma(\mathcal{C}) = \int_{\lambda=0}^1 D_F(\mathbf{p}_\lambda || \mathbf{p}_0) d\lambda$. Fix some inner point $\mathbf{p} \in \Gamma_F(\mathbf{p}_0, \mathbf{p}_1) \setminus \{\mathbf{p}_0, \mathbf{p}_1\}$. From the three-point

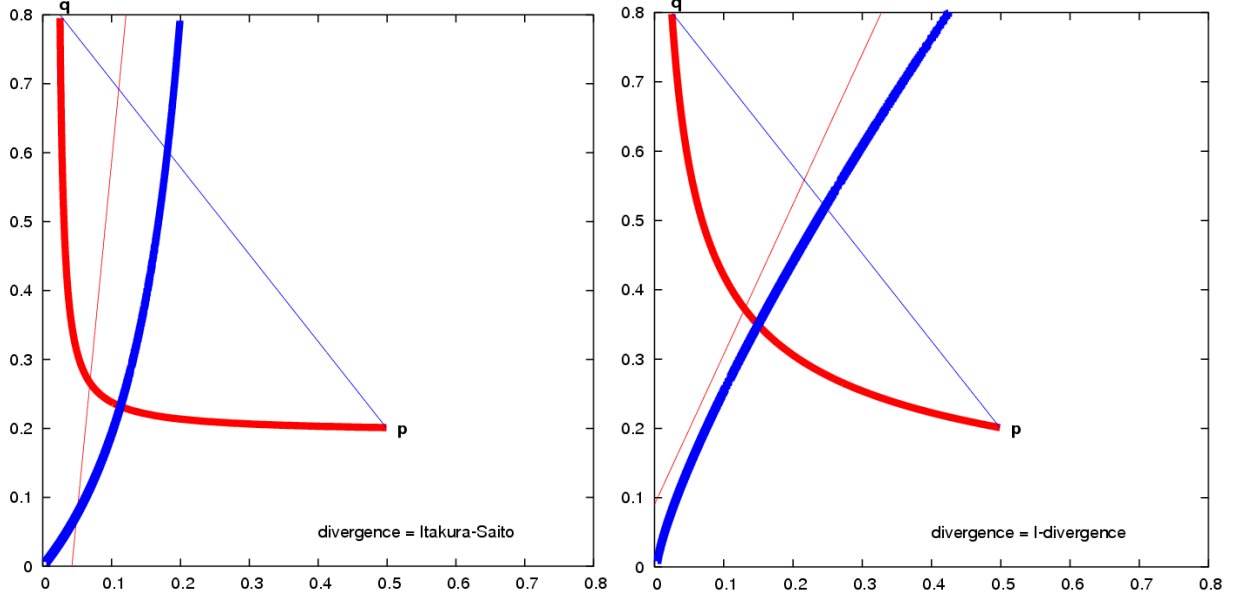


Figure 8: Bregman bisectors and their relationships with respect to $\Lambda(\mathbf{p}, \mathbf{q})$ (straight line segments) and $\Gamma_F(\mathbf{p}, \mathbf{q})$ (bold curves), for the Itakura-Saito divergence (left) and I-divergence (right). Bold curves become linear in \mathcal{X}' ; colors depict the Bregman orthogonality relationships of Lemma 8.

property (Property 5), the set of points $\{\mathbf{y} \in \mathcal{X} \mid D_F(\mathbf{y} \parallel \mathbf{p}_0) = D_F(\mathbf{y} \parallel \mathbf{p}) + D_F(\mathbf{p} \parallel \mathbf{p}_0)\}$ is the hyperplane $H_{\mathbf{p}} : \langle \mathbf{y}, \mathbf{h} \rangle = b$ (\mathbf{h} is a perpendicular vector to the hyperplane) which splits \mathcal{X} into two open half-spaces $H_{\mathbf{p}}^+ : \langle \mathbf{y}, \mathbf{h} \rangle > b$, and $H_{\mathbf{p}}^- : \langle \mathbf{y}, \mathbf{h} \rangle < b$. Now, $H_{\mathbf{p}}$ intersects $\Gamma(\mathbf{p}_0, \mathbf{p}_1)$ since $H_{\mathbf{p}}$ separates \mathbf{p}_0 from \mathbf{p}_1 . Indeed, $H_{\mathbf{p}}(\mathbf{p}_0) = \langle \mathbf{p}_0 - \mathbf{p}, \mathbf{p}' - \mathbf{p}' \rangle = D_F(\mathbf{p}_0 \parallel \mathbf{p}) + D_F(\mathbf{p} \parallel \mathbf{p}_0) > 0$ and $H_{\mathbf{p}}(\mathbf{p}_1) = \langle \mathbf{p}_1 - \mathbf{p}, \mathbf{p}'_0 - \mathbf{p}' \rangle = \frac{\lambda-1}{\lambda} \langle \mathbf{p}_1 - \mathbf{p}, \mathbf{p}'_1 - \mathbf{p}' \rangle < 0$ where $\mathbf{p}' = \lambda \mathbf{p}'_0 + (1 - \lambda) \mathbf{p}'_1$ (with $\lambda \in]0, 1[$). Therefore any connected path \mathcal{C} joining \mathbf{p}_0 to \mathbf{p}_1 has to intersect $H_{\mathbf{p}}$.

To finish up, consider function $f : [0, 1] \rightarrow \mathcal{C}$ with $f(0) = \mathbf{p}_0$, $f(1) = \mathbf{p}_1$, and $f(\lambda) \in \mathcal{C} \cap H_{\mathbf{p}_\lambda}$ otherwise, where it is understood that \mathbf{p}_λ is hereafter a point of $\Gamma_F(\mathbf{p}_0, \mathbf{p}_1)$. Since $f(\lambda) \in H_{\mathbf{p}(\lambda)}$, we have $D_F(f(\lambda) \parallel \mathbf{p}_0) = D_F(f(\lambda) \parallel \mathbf{p}_\lambda) + D_F(\mathbf{p}_\lambda \parallel \mathbf{p}_0) \geq D_F(\mathbf{p}_\lambda \parallel \mathbf{p}_0)$, with equality if and only if $f(\lambda) = \mathbf{p}_\lambda$. Thus we have

$$\ell_{\Gamma}(\Gamma_F(\mathbf{p}_0, \mathbf{p}_1)) = \int_{\lambda=0}^1 D_F(\mathbf{p}_\lambda \parallel \mathbf{p}_0) d\lambda \leq \int_{\lambda=0}^1 D_F(f(\lambda) \parallel \mathbf{p}_0) d\lambda \leq \ell_{\Gamma}(\mathcal{C}) .$$

The case of $\Lambda(\mathbf{p}_0, \mathbf{p}_1)$ follows similarly from Legendre convex duality.

□

Corollary 1 *Since $\Gamma_F(\mathbf{p}_0, \mathbf{p}_1) = \Gamma_F(\mathbf{p}_1, \mathbf{p}_0)$ (respectively, since $\Lambda(\mathbf{p}_0, \mathbf{p}_1) = \Lambda(\mathbf{p}_1, \mathbf{p}_0)$) we deduce that $\Gamma_F(\mathbf{p}_0, \mathbf{p}_1)$ minimizes also $\int_{\lambda=0}^1 D_F(\mathbf{p}_1 \parallel \mathbf{p}_\lambda) d\lambda$ (respectively, minimizes also $\int_{\lambda=0}^1 D_F(\mathbf{p}_\lambda \parallel \mathbf{p}_1) d\lambda$) over all curves $\mathcal{C} = \{\mathbf{p}_\lambda\}_{\lambda=0}^1$.*

Observe also that $\Gamma_F(\mathbf{p}, \mathbf{q})$ is the unique geodesic path joining \mathbf{p} to \mathbf{q} in \mathcal{X} for the metric image by $\nabla^{-1}F$ of the Euclidean metric.

Finally, we give a characterization of these geodesics in information-theoretic spaces. Recall that Banerjee et al. [6] showed that Bregman divergences are in bijection with exponential families. This was emphasized by Theorem 2 that proved that the Kullback-Leibler divergence of probability density functions of the same exponential family \mathcal{E}_F is a Bregman divergence D_F for the cumulant function F . From this standpoint, $\Lambda(\mathbf{p}, \mathbf{q})$ and $\Gamma_F(\mathbf{p}, \mathbf{q})$ minimize the total Kullback-Leibler divergence, a characteristic that we choose to call the *information length* of a curve. Since the Kullback-Leibler divergence is not symmetric, this justifies for the existence of two geodesics, one which appears to be linear when parameterized with the natural affine coordinate system $(\boldsymbol{\theta})$, and the other that is linear in the expectation affine coordinate system $(\boldsymbol{\mu})$. See also [1].

Corollary 2 *Suppose $p(\cdot|\boldsymbol{\theta}_0)$ and $p(\cdot|\boldsymbol{\theta}_1)$ are probability density functions of the same exponential family \mathcal{E}_F . Then $\Gamma_F(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$ (resp. $\Lambda(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$) minimizes $\ell_\Gamma(\mathcal{C}) = \int_{\lambda=0}^1 \text{KL}(\boldsymbol{\theta}_0||\boldsymbol{\theta}_\lambda)d\lambda$ (resp. $\ell_\Lambda(\mathcal{C}) = \int_{\lambda=0}^1 \text{KL}(\boldsymbol{\theta}_\lambda||\boldsymbol{\theta}_0)d\lambda$) over all curves $\mathcal{C} = \{p(\cdot|\boldsymbol{\theta}_\lambda)\}_{\lambda=0}^1$.*

4 Bregman Voronoi diagrams

Let $\mathcal{S} = \{\mathbf{p}_1, \dots, \mathbf{p}_n\}$ be a finite point set of $\mathcal{X} \subset \mathbb{R}^d$. To each point \mathbf{p}_i is attached a d -variate continuous function D_i defined over \mathcal{X} . We define the *lower envelope* of the functions as the graph of $\min_{1 \leq i \leq n} D_i$ and their *minimization diagram* as the subdivision of \mathcal{X} into cells such that, in each cell, $\arg \min_i f_i$ is fixed.

The Euclidean Voronoi diagram is the minimization diagram for $D_i(\mathbf{x}) = \|\mathbf{x} - \mathbf{p}_i\|^2$. In this section, we introduce Bregman Voronoi diagrams as minimization diagrams of Bregman divergences (see Figure 10).

We define three types of Bregman Voronoi diagrams in §4.1. We establish a correspondence between Bregman Voronoi diagrams and polytopes in §4.2 and with power diagrams in §4.3. These correspondences lead to tight combinatorial bounds and efficient algorithms. Finally, in §4.4, we give two generalizations of Bregman Voronoi diagrams; k -order and k -bag diagrams.

We note $\mathcal{S}' = \{\nabla_F(\mathbf{p}_i), i = 1, \dots, n\}$ the *gradient point set* associated to \mathcal{S} .

4.1 Three types of diagrams

Because Bregman divergences are not necessarily symmetric, we associate to each site \mathbf{p}_i *two types* of distance functions, namely $D_i(\mathbf{x}) = D_F(\mathbf{x}||\mathbf{p}_i)$ and $D'_i(\mathbf{x}) = D_F(\mathbf{p}_i||\mathbf{x})$. The *minimization diagram* of the D_i , $i = 1, \dots, n$, is called the *first-type* Bregman Voronoi

diagram of \mathcal{S} , which we denote by $\text{vor}_F(\mathcal{S})$. The d -dimensional cells of this diagram are in *1-1 correspondence* with the sites \mathbf{p}_i and the d -dimensional cell of \mathbf{p}_i is defined as

$$\text{vor}_F(\mathbf{p}_i) \stackrel{\text{def}}{=} \{\mathbf{x} \in \mathcal{X} \mid D_F(\mathbf{x} \parallel \mathbf{p}_i) \leq D_F(\mathbf{x} \parallel \mathbf{p}_j) \ \forall \mathbf{p}_j \in \mathcal{S}\}$$

Since the Bregman bisectors of the first-type are hyperplanes, the cells of any diagram of the first-type are convex polyhedra. Therefore, first-type Bregman Voronoi diagrams are *affine* diagrams [4, 5].

Similarly, the minimization diagram of the D'_i , $i = 1, \dots, n$, is called the *second-type* Bregman Voronoi diagram of \mathcal{S} , which we denote by $\text{vor}'_F(\mathcal{S})$. A cell in $\text{vor}'_F(\mathcal{S})$ is associated to each site \mathbf{p}_i and is defined as above with permuted divergence arguments:

$$\text{vor}'_F(\mathbf{p}_i) \stackrel{\text{def}}{=} \{\mathbf{x} \in \mathcal{X} \mid D_F(\mathbf{p}_i \parallel \mathbf{x}) \leq D_F(\mathbf{p}_j \parallel \mathbf{x}) \ \forall \mathbf{p}_j \in \mathcal{S}\}$$

In contrast with the diagrams of the first-type, the diagrams of the second type have, in general, curved faces.

Figure 9 illustrates these Bregman Voronoi diagrams for the Kullback-Leibler and the Itakura-Saito divergences. Note that the Euclidean Voronoi diagram is a Bregman Voronoi diagram since $\text{vor}(\mathcal{S}) = \text{vor}_F(\mathcal{S}) = \text{vor}'_F(\mathcal{S})$ for $F(\mathbf{x}) = \|\mathbf{x}\|^2$.

For asymmetric Bregman divergences D_F , we can further consider the symmetrized Bregman divergence $S_F = D_{\tilde{F}}$ and define a *third-type* Bregman Voronoi diagram $\text{vor}''_F(\mathcal{S})$. The cell of $\text{vor}''_F(\mathcal{S})$ associated to site \mathbf{p}_i is defined as:

$$\text{vor}''_F(\mathbf{p}_i) \stackrel{\text{def}}{=} \{\mathbf{x} \in \mathcal{X} \mid S_F(\mathbf{x}, \mathbf{p}_i) \leq S_F(\mathbf{x}, \mathbf{p}_j) \ \forall \mathbf{p}_j \in \mathcal{S}\}$$

From the Legendre duality between divergences, we deduce correspondences between the diagrams of the first and the second types. As usual, F^* is the convex conjugate of F .

Lemma 10 $\text{vor}'_F(\mathcal{S}) = \nabla^{-1} F(\text{vor}_{F^*}(\mathcal{S}'))$ and $\text{vor}_F(\mathcal{S}) = \nabla^{-1} F(\text{vor}'_{F^*}(\mathcal{S}'))$.

Proof: By Lemma 3, we have $D_F(\mathbf{x} \parallel \mathbf{y}) = D_{F^*}(\mathbf{y}' \parallel \mathbf{x}')$, which gives $\text{vor}_F(\mathbf{p}_i) = \{\mathbf{x} \in \mathcal{X} \mid D_{F^*}(\mathbf{p}'_i \parallel \mathbf{x}') \leq D_{F^*}(\mathbf{p}'_j \parallel \mathbf{x}') \ \forall \mathbf{p}'_j \in \mathcal{S}'\} = \nabla^{-1} F(\text{vor}'_{F^*}(\mathbf{p}'_i))$. The proof of the second part follows the same path. \square

Hence, constructing the second-type curved diagram $\text{vor}'_F(\mathcal{S})$ reduces to constructing an affine diagram in the gradient space \mathcal{X}' (and map the cells by ∇F^{-1}).

Let us end this section by considering the case of symmetrized Bregman divergences introduced in §2.3: $S_F(\mathbf{p}, \mathbf{q}) = D_{\tilde{F}}(\tilde{\mathbf{p}} \parallel \tilde{\mathbf{q}}) = D_{\tilde{F}}(\tilde{\mathbf{q}} \parallel \tilde{\mathbf{p}})$ where \tilde{F} is a $2d$ -variate function and $\tilde{\mathbf{x}} = [\mathbf{x} \ \mathbf{x}']^T$. As already noted in §2.3, $\tilde{\mathbf{x}}$ lies on the d -manifold $\tilde{\mathcal{X}} = \{[\mathbf{x} \ \mathbf{x}']^T \mid \mathbf{x} \in \mathbb{R}^d\}$. It follows that the symmetrized Voronoi diagram $\text{vor}''_F(\mathcal{S})$ is the projection of the restriction to $\tilde{\mathcal{X}}$ of the affine diagram $\text{vor}_{\tilde{F}}(\tilde{\mathcal{S}})$ of \mathbb{R}^{2d} where $\tilde{\mathcal{S}} = \{\tilde{\mathbf{p}}_i, \mathbf{p}_i \in \mathcal{S}\}$. Hence, computing the symmetrized Voronoi diagram of \mathcal{S} reduces to:

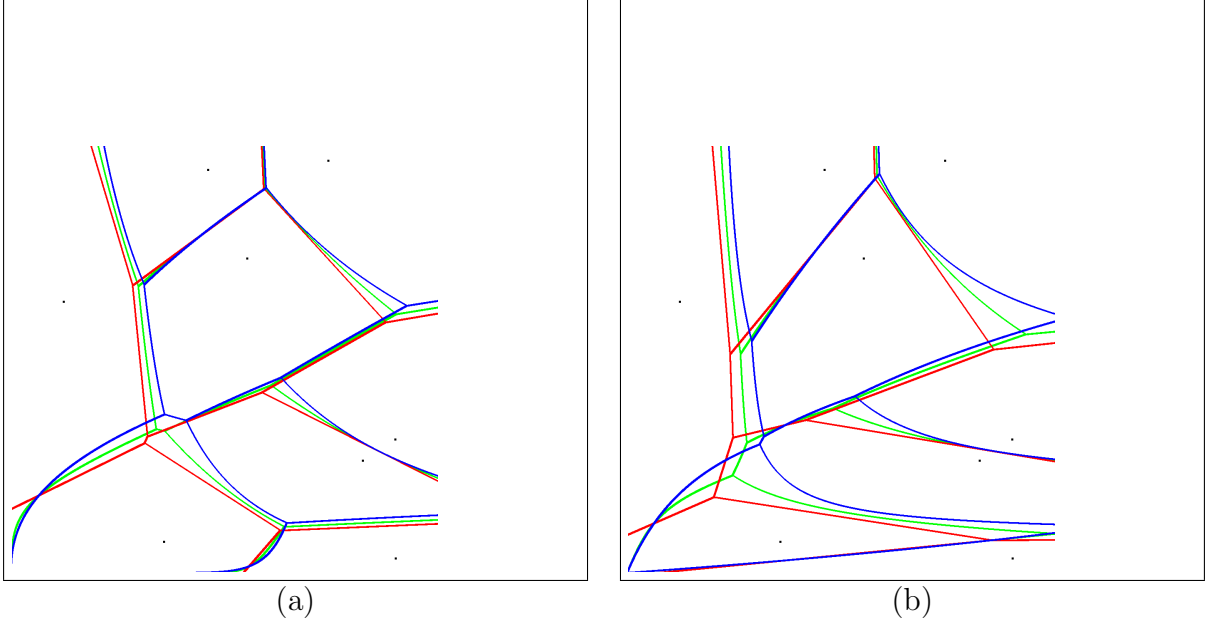


Figure 9: Three types of Bregman Voronoi diagrams for (a) the Kullback-Leibler and (b) the Itakura-Saito divergences. First-type affine Bregman Voronoi diagram (red), second-type Bregman Voronoi diagram (blue) and symmetrized Bregman Voronoi diagram (green).

1. computing the first-type Bregman Voronoi diagram $\text{vor}_{\tilde{F}}(\tilde{\mathcal{S}})$ of \mathbb{R}^{2d} ,
2. intersecting the cells of this diagram with the manifold $\tilde{\mathcal{X}}$, and
3. projecting all points of $\text{vor}_{\tilde{F}}(\tilde{\mathcal{S}}) \cap \tilde{\mathcal{X}}$ to \mathcal{X} by simply dropping the last d coordinates.

4.2 Bregman Voronoi diagrams from polytopes

Let $H_{\mathbf{p}_i}$, $i = 1, \dots, n$, denote the hyperplanes of $\hat{\mathcal{X}}$ defined in §3.2. For any $\mathbf{x} \in \mathcal{X}$, we have following Lemma 1

$$D_F(\mathbf{x}||\mathbf{p}_i) \leq D_F(\mathbf{x}||\mathbf{p}_j) \iff H_{\mathbf{p}_i}(\mathbf{x}) \geq H_{\mathbf{p}_j}(\mathbf{x}).$$

The first-type Bregman Voronoi diagram of \mathcal{S} is therefore the maximization diagram of the n linear functions $H_{\mathbf{p}_i}(\mathbf{x})$ whose graphs are the hyperplanes $H_{\mathbf{p}_i}$ (see Figure 10). Equivalently, we have

Theorem 6 *The first-type Bregman Voronoi diagram $\text{vor}_F(\mathcal{S})$ is obtained by projecting by $\text{Proj}_{\mathcal{X}}$ the faces of the $(d+1)$ -dimensional convex polyhedron $\mathcal{H} = \cap_i H_{\mathbf{p}_i}^\uparrow$ of \mathcal{X}^+ onto \mathcal{X} .*

From McMullen’s upperbound theorem [31] and Chazelle’s optimal half-space intersection algorithm [14], we know that the intersection of n halfspaces of \mathbb{R}^d has complexity $\Theta(n^{\lfloor \frac{d}{2} \rfloor})$

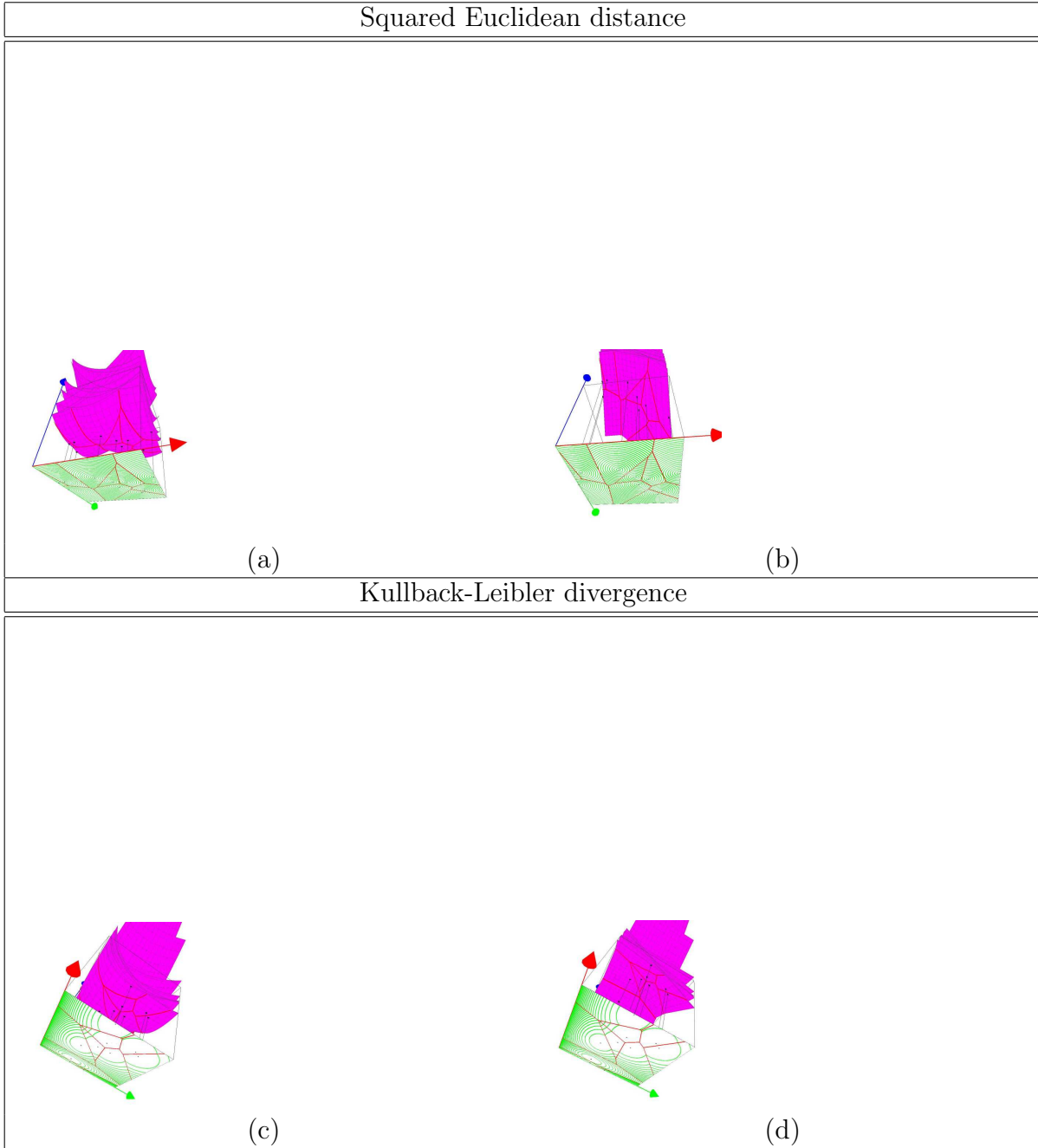


Figure 10: Voronoi diagrams as minimization diagrams. The first row shows minimization diagrams for the Euclidean distance and the second row shows minimization diagrams for the Kullback-Leibler divergence. In the first column, the functions are the non-linear functions $D_i(\mathbf{x})$ and, in the second column, the functions are the linear functions $H_{\mathbf{p}_i}(\mathbf{x})$, both leading to the same minimization diagrams. Isolines are shown in green.

and can be computed in optimal-time $\Theta(n \log n + n^{\lfloor \frac{d}{2} \rfloor})$ for any fixed dimension d . From Theorem 6 and Lemma 10, we then deduce the following theorem.

Theorem 7 *The Bregman Voronoi diagrams of type 1 or 2 of a set of n d -dimensional points have complexity $\Theta(n^{\lfloor \frac{d+1}{2} \rfloor})$ and can be computed in optimal time $\Theta(n \log n + n^{\lfloor \frac{d+1}{2} \rfloor})$. The third-type Bregman Voronoi diagram for the symmetrized Bregman divergence of a set of n d -dimensional points has complexity $O(n^d)$ and can be obtained in time $O(n^d)$.*

Apart from Chazelle's algorithm, several other algorithms are known for constructing the intersection of a finite number of halfplanes, especially in the 2- and 3-dimensional cases. See [10, 5] for further references.

4.3 Bregman Voronoi diagrams from power diagrams

The power distance of a point \mathbf{x} to a Euclidean ball $B = B(\mathbf{p}, r)$ is defined as $\|\mathbf{p} - \mathbf{x}\|^2 - r^2$. Given n balls $B_i = B(\mathbf{p}_i, r_i)$, $i = 1, \dots, n$, the *power diagram* (or Laguerre diagram) of the B_i is defined as the minimization diagram of the corresponding n functions $D_i(\mathbf{x}) = \|\mathbf{p}_i - \mathbf{x}\|^2 - r_i^2$. The power bisector of any two balls $B(\mathbf{p}_i, r_i)$ and $B(\mathbf{p}_j, r_j)$ is the radical hyperplane of equation $2\langle \mathbf{x}, \mathbf{p}_j - \mathbf{p}_i \rangle + \|\mathbf{p}_i\|^2 - \|\mathbf{p}_j\|^2 + r_j^2 - r_i^2 = 0$. Thus power diagrams are affine diagrams. In fact, as shown by Aurenhammer [3, 10], *any* affine diagram is *identical* to the power diagram of a set of corresponding balls. In general, some balls may have an empty cell in their power diagram.

Since Bregman Voronoi diagrams of the first type are affine diagrams, Bregman Voronoi diagrams are power diagrams [3, 10] in disguise. The following theorem makes precise the correspondence between Bregman Voronoi diagrams and power diagrams (see Figure 11).

Theorem 8 *The first-type Bregman Voronoi diagram of n sites is identical to the power diagram of the n Euclidean spheres of equations*

$$\langle \mathbf{x} - \mathbf{p}'_i, \mathbf{x} - \mathbf{p}'_i \rangle = \langle \mathbf{p}'_i, \mathbf{p}'_i \rangle + 2(F(\mathbf{p}_i) - \langle \mathbf{p}_i, \mathbf{p}'_i \rangle), \quad i = 1, \dots, n.$$

Proof: We have

$$\begin{aligned} D_F(\mathbf{x} | \mathbf{p}_i) &\leq D_F(\mathbf{x} | \mathbf{p}_j) \\ \iff -F(\mathbf{p}_i) - \langle \mathbf{x} - \mathbf{p}_i, \mathbf{p}'_i \rangle &\leq -F(\mathbf{p}_j) - \langle \mathbf{x} - \mathbf{p}_j, \mathbf{p}'_j \rangle \end{aligned}$$

Multiplying twice the last inequality, and adding $\langle \mathbf{x}, \mathbf{x} \rangle$ to both sides yields

$$\begin{aligned} \langle \mathbf{x}, \mathbf{x} \rangle - 2\langle \mathbf{x}, \mathbf{p}'_i \rangle - 2F(\mathbf{p}_i) + 2\langle \mathbf{p}_i, \mathbf{p}'_i \rangle &\leq \langle \mathbf{x}, \mathbf{x} \rangle - 2\langle \mathbf{x}, \mathbf{p}'_j \rangle - 2F(\mathbf{p}_j) + 2\langle \mathbf{p}_j, \mathbf{p}'_j \rangle \\ \iff \langle \mathbf{x} - \mathbf{p}'_i, \mathbf{x} - \mathbf{p}'_i \rangle - r_i^2 &\leq \langle \mathbf{x} - \mathbf{p}'_j, \mathbf{x} - \mathbf{p}'_j \rangle - r_j^2, \end{aligned}$$

where $r_i^2 = \langle \mathbf{p}'_i, \mathbf{p}'_i \rangle + 2(F(\mathbf{p}_i) - \langle \mathbf{p}_i, \mathbf{p}'_i \rangle)$ and $r_j^2 = \langle \mathbf{p}'_j, \mathbf{p}'_j \rangle + 2(F(\mathbf{p}_j) - \langle \mathbf{p}_j, \mathbf{p}'_j \rangle)$. The last inequality means that the power of \mathbf{x} with respect to the Euclidean (possibly imaginary) ball $B(\mathbf{p}'_i, r_i)$ is no more than the power of \mathbf{x} with respect to the Euclidean (possibly imaginary) ball $B(\mathbf{p}'_j, r_j)$. \square

As already noted, for $F(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$, $\text{vor}_F(\mathcal{S})$ is the Euclidean Voronoi diagram of \mathcal{S} . Accordingly, the theorem says that the centers of the spheres are the \mathbf{p}_i and $r_i^2 = 0$ since $\mathbf{p}'_i = \mathbf{p}_i$. Figure 11 displays affine Bregman Voronoi diagrams³ and their equivalent power diagrams for the squared Euclidean, Kullback-Leibler and exponential divergences.

Note that although the affine Bregman Voronoi diagram obtained by scaling the divergence D_F by a factor $\lambda > 0$ does not change, the equivalent power diagrams are not *strictus sensu* identical since the centers of corresponding Euclidean balls and radii are mapped differently. See the example of the squared Euclidean distance depicted in Figure 11(a). Since Power diagrams are well defined “everywhere”, this equivalence relationship provides a natural way to extend the scope of definition of Bregman Voronoi diagrams from $\mathcal{X} \subset \mathbb{R}^d$ to the full space \mathbb{R}^d . (That is, Bregman Voronoi diagrams are power diagrams restricted to \mathcal{X} .)

To check that associated balls may be potentially imaginary, consider for example, the Kullback-Leibler divergence. The Bregman generator function is $F(\mathbf{x}) = \sum_i x_i \log x_i$ and the gradient is $\nabla F(\mathbf{x}) = [\log x_1 \dots \log x_d]^T$. A point $\mathbf{p} = [p_1 \dots p_d]^T \in \mathcal{X}$ maps to a Euclidean ball of center $\mathbf{p}' = [\log p_1 \dots \log p_d]^T$ with radius $r_{\mathbf{p}}^2 = \sum_i (\log^2 p_i - 2p_i)$. Thus for points \mathbf{p} with coordinates $p_i > \frac{1}{2} \log p_i^2$ for $i \in \{1, \dots, d\}$, the squared radius $r_{\mathbf{p}}^2$ is negative, yielding an imaginary ball. See Figure 11(b).

It is also to be observed that not all power diagrams are Bregman Voronoi diagrams. Indeed, in power diagrams, some balls may have empty cells while each site has necessarily a non empty cell in a Bregman Voronoi diagram (See Figure 11 and Section 4.4 for a further discussion at this point).

Since there exist fast algorithms for constructing power diagrams [36], Theorem 8 provides an *efficient* way to construct Bregman Voronoi diagrams.

4.4 Generalized Bregman divergences and their Voronoi diagrams

Weighted Bregman Voronoi diagrams

Let us associate to each site \mathbf{p}_i a weight $w_i \in \mathbb{R}$. We define the *weighted divergence* between two weighted points as $WD_F(\mathbf{p}_i || \mathbf{p}_j) \stackrel{\text{def}}{=} D_F(\mathbf{p}_i || \mathbf{p}_j) + w_i - w_j$. We can define bisectors and weighted Bregman Voronoi diagrams in very much the same way as for non weighted divergences. The Bregman Voronoi region associated to the weighted point (\mathbf{p}_i, w_i) is defined as

$$\text{vor}_F(\mathbf{p}_i, w_i) = \{\mathbf{x} \in \mathcal{X} \mid D_F(\mathbf{x} || \mathbf{p}_i) + w_i \leq D_F(\mathbf{x} || \mathbf{p}_j) + w_j \forall \mathbf{p}_j \in \mathcal{S}\}.$$

³See JavaTM applet at <http://www.cs1.sony.co.jp/person/nielsen/BVDapplet/>

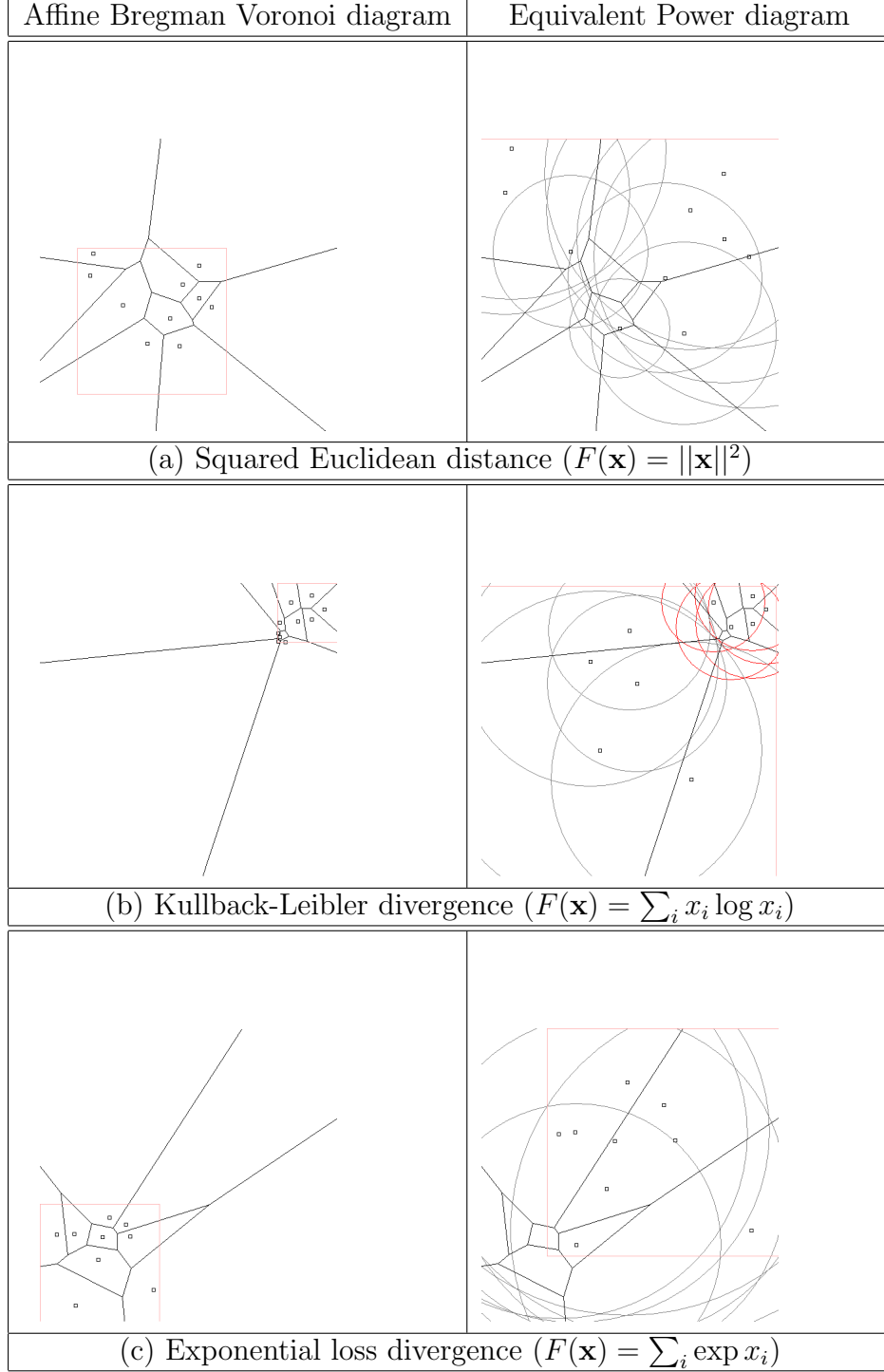


Figure 11: Affine Bregman Voronoi diagrams (left column) can be computed as power diagrams (right column). Illustrations for the squared Euclidean distance (a), Kullback-Leibler divergence (b), and exponential divergence (c). Circles are drawn either in grey to denote positive radii, or in red to emphasize imaginary radii. Observe that although some cells of the power diagrams may be empty, all cells of the affine Bregman Voronoi diagram are necessarily non-empty.

Observe that the bisectors of the first-type diagrams are still hyperplanes and that the diagram can be obtained as the projection of a convex polyhedron or as the power diagram of a finite set of balls. The only difference with respect to the construction of Section 4.2 is the fact that now the hyperplanes $H_{\mathbf{p}_i}$ are no longer tangent to \mathcal{F} since they are *shifted* by a z -displacement of length w_i . Hence Theorem 7 extends to weighted Bregman Voronoi diagrams.

Theorem 9 *The weighted Bregman Voronoi diagrams of type 1 or 2 of a set of n d -dimensional points have complexity $\Theta(n^{\lfloor \frac{d+1}{2} \rfloor})$ and can be computed in optimal time $\Theta(n \log n + n^{\lfloor \frac{d+1}{2} \rfloor})$.*

k -order Bregman Voronoi diagrams

We define the k -order Bregman Voronoi diagram of n punctual sites of \mathcal{X} as the subdivision of \mathcal{X} into cells such that each cell is associated to a subset $\mathcal{T} \subset \mathcal{S}$ of k sites and consists of the points of \mathcal{X} whose divergence to any site in \mathcal{T} is less than the divergence to the sites not in \mathcal{T} . Similarly to the case of higher-order Euclidean Voronoi diagrams, we have:

Theorem 10 *The k -order Bregman Voronoi diagram of n d -dimensional points is a weighted Bregman Voronoi diagram.*

Proof: Let $\mathcal{S}_1, \mathcal{S}_2, \dots$ denote the subsets of k points of \mathcal{S} and write

$$\begin{aligned} D_i(\mathbf{x}) &= \frac{1}{k} \sum_{\mathbf{p}_j \in \mathcal{S}_i} D_F(\mathbf{x} || \mathbf{p}_j) \\ &= F(\mathbf{x}) - \frac{1}{k} \sum_{\mathbf{p}_j \in \mathcal{S}_i} F(\mathbf{p}_j) + \frac{1}{k} \sum_{\mathbf{p}_j \in \mathcal{S}_i} \langle \mathbf{x} - \mathbf{p}_j, \mathbf{p}'_j \rangle \\ &= F(\mathbf{x}) - F(\mathbf{c}_i) - \langle \mathbf{x} - \mathbf{c}_i, \mathbf{c}'_i \rangle + w_i \\ &= WD_F(\mathbf{x} || \mathbf{c}_i) \end{aligned}$$

where $\mathbf{c}_i = \nabla^{-1} F \left(\frac{1}{k} \sum_{j \in \mathcal{S}_i} \mathbf{p}'_j \right)$ and the weight associated to \mathbf{c}_i is $w_i = F(\mathbf{c}_i) - \langle \mathbf{c}_i, \mathbf{c}'_i \rangle - \frac{1}{k} \sum_{j \in \mathcal{S}_i} (F(\mathbf{p}_j) + \langle \mathbf{p}_j, \mathbf{p}'_j \rangle)$.

Hence, \mathcal{S}_i is the set of the k nearest neighbors of \mathbf{x} iff $D_i(\mathbf{x}) \leq D_j(\mathbf{x})$ for all j or, equivalently, iff \mathbf{x} belongs to the cell of \mathbf{c}_i in the weighted Bregman Voronoi diagram of the \mathbf{c}_i . \square

k -bag Bregman Voronoi diagrams

Let F_1, \dots, F_k be k strictly convex and differentiable functions, and $\boldsymbol{\alpha} = [\alpha_1 \dots \alpha_k]^T \in \mathbb{R}_+^k$ a vector of positive weights. Consider the d -variate function $F_{\boldsymbol{\alpha}} = \sum_{l=1}^k \alpha_l F_l$. By virtue of the

positive additivity property rule of Bregman basis functions (Property 3), D_{F_α} is a Bregman divergence.

Now consider a set $\mathcal{S} = \{\mathbf{p}_1, \dots, \mathbf{p}_n\}$ of n points of \mathbb{R}^d . To each site \mathbf{p}_i , we associate a weight vector $\alpha_i = [\alpha_i^{(1)} \dots \alpha_i^{(k)}]^T$ inducing a Bregman divergence $D_{F_{\alpha_i}}(\mathbf{x}||\mathbf{p}_i) \stackrel{\text{def}}{=} D_{\alpha_i}(\mathbf{x}||\mathbf{p}_i)$ anchored at that site. Let us consider the first-type of k -bag Bregman Voronoi diagram (k -bag BVD for short). The first-type bisector $K_F(\mathbf{p}_i, \mathbf{p}_j)$ of two weighted points (\mathbf{p}_i, α_i) and (\mathbf{p}_j, α_j) is the locus of points \mathbf{x} at equidivergence to \mathbf{p}_i and \mathbf{p}_j . That is, $K_F(\mathbf{p}_i, \mathbf{p}_j) = \{\mathbf{x} \in \mathcal{X} \mid D_{\alpha_i}(\mathbf{x}||\mathbf{p}_i) = D_{\alpha_j}(\mathbf{x}||\mathbf{p}_j)\}$. The equation of the bisector is simply obtained using the definition of Bregman divergences (Eq. 1) as

$$F_{\alpha_i}(\mathbf{x}) - F_{\alpha_i}(\mathbf{p}_i) - \langle \mathbf{x} - \mathbf{p}_i, \nabla F_{\alpha_i}(\mathbf{p}_i) \rangle = F_{\alpha_j}(\mathbf{x}) - F_{\alpha_j}(\mathbf{p}_j) - \langle \mathbf{x} - \mathbf{p}_j, \nabla F_{\alpha_j}(\mathbf{p}_j) \rangle.$$

This yields the equation of the first-type bisector $K_F(\mathbf{p}_i, \mathbf{p}_j)$

$$\sum_{l=1}^k (\alpha_i^{(l)} - \alpha_j^{(l)}) F_l(\mathbf{x}) - \langle \mathbf{x}, \nabla F_{\alpha_j}(\mathbf{p}_j) - \nabla F_{\alpha_i}(\mathbf{p}_i) \rangle + c = 0, \quad (13)$$

where c is a *constant* depending on weighted sites (\mathbf{p}_i, α_i) and (\mathbf{p}_j, α_j) . Note that the equation of the first-type k -bag BVD bisector is linear if and only if $\alpha_i = \alpha_j$ (i.e., the case of standard BVDs).

Let us consider the linearization lifting $\mathbf{x} \mapsto \hat{\mathbf{x}} = [\mathbf{x} \ F_1(\mathbf{x}) \dots F_k(\mathbf{x})]^T$ that maps a point $\mathbf{x} \in \mathbb{R}^d$ into a point $\hat{\mathbf{x}}$ in \mathbb{R}^{d+k} . Then Eq. 13 becomes linear, namely $\langle \hat{\mathbf{x}}, \mathbf{a} \rangle + c = 0$ with

$$\mathbf{a} = \begin{bmatrix} \nabla F_{\alpha_j}(\mathbf{p}_j) - \nabla F_{\alpha_i}(\mathbf{p}_i) \\ \alpha_i - \alpha_j \end{bmatrix} \in \mathbb{R}^{d+k}.$$

That is, first-type bisectors of a k -bag BVD are hyperplanes of \mathbb{R}^{d+k} . Therefore the complexity of a k -bag Voronoi diagram is at most $O(n^{\lfloor \frac{k+d}{2} \rfloor})$, since it can be obtained as the intersection of the affine Voronoi diagram in \mathbb{R}^{d+k} with the convex d -dimensional submanifold $\{\hat{\mathbf{x}} = [\mathbf{x} \ F_1(\mathbf{x}) \dots F_k(\mathbf{x})]^T \mid \mathbf{x} \in \mathbb{R}^d\}$.

Theorem 11 *The k -bag Voronoi diagram (for $k > 1$) on a bag of d -variate Bregman divergences of a set of n points of \mathbb{R}^d has combinatorial complexity $O(n^{\lfloor \frac{k+d}{2} \rfloor})$ and can be computed within the same time bound.*

Further, using the Legendre transform, we define a second-type (dual) k -bag BVD. We have $\nabla F_\alpha = \sum_{l=1}^k \alpha_l \nabla F_l$ and $F_\alpha^* = \int \nabla F_\alpha^{-1}$. (Observe that $F_\alpha^* \neq \sum_{l=1}^k \alpha_l F_l^*$ in general.)

k -bag Bregman Voronoi diagrams are closely related to the anisotropic diagrams of Labelle and Shewchuk [27] that associate to *each* point $\mathbf{x} \in \mathcal{X}$ a metric tensor $\mathbf{M}_\mathbf{x}$ which tells how lengths and angles should be measured from the local perspective of \mathbf{x} . Labelle and Shewchuk relies on a deformation tensor (ideally defined everywhere) to compute the distance between

any two points \mathbf{p} and \mathbf{q} from the perspective of \mathbf{x} as $d_{\mathbf{x}}(\mathbf{p}, \mathbf{q}) = \sqrt{(\mathbf{p} - \mathbf{q})^T \mathbf{M}_{\mathbf{x}} (\mathbf{p} - \mathbf{q})}$. Let $d_{\mathbf{x}}(\mathbf{p}) = d_{\mathbf{x}}(\mathbf{x}, \mathbf{p})$. The anisotropic Voronoi diagram, which approximates the ideal but computationally prohibitive Riemannian Voronoi diagram, is defined as the arrangement of the following anisotropic Voronoi cells:

$$\text{Vor}(\mathbf{p}_i) = \{\mathbf{x} \in \mathcal{X} \mid d_{\mathbf{p}_i}(\mathbf{x}) \leq d_{\mathbf{p}_j}(\mathbf{x}) \ \forall j \in \{1, \dots, n\}\}, \ \forall i \in \{1, \dots, n\}.$$

It follows that all anisotropic Voronoi cells are non-empty as it is the case for k -bag Bregman Voronoi diagrams.

Hence, the site weights of a k -bag Bregman Voronoi diagram sparsely define a *tensor divergence* that indicates how divergences should be measured locally from the respective bag of divergences. Noteworthy, our study of k -bag Bregman Voronoi diagrams shows that the anisotropic Voronoi diagram also admits a *second-type* anisotropic Voronoi diagram, induced by the respective dual Legendre functions of the Bregman basis functions of the quadratic distance monomials. The Legendre dual of a quadratic distance function $d_{\mathbf{M}}(\mathbf{p}, \mathbf{q}) = (\mathbf{p} - \mathbf{q})^T \mathbf{M} (\mathbf{p} - \mathbf{q})$ induced by positive-definite matrix \mathbf{M} is the quadratic distance $d_{\mathbf{M}^{-1}}$. (Matrix \mathbf{M} is itself usually obtained as the inverse of a variance-covariance matrix $\mathbf{\Sigma}$ in so-called Mahalanobis distances.)

5 Bregman triangulations

Consider the Euclidean Voronoi diagram $\text{vor}(\mathcal{S})$ of a finite set \mathcal{S} of points of \mathbb{R}^d (called sites). Let f be a face of $\text{vor}(\mathcal{S})$ that is the intersection of k d -cells of $\text{vor}(\mathcal{S})$. We associate to f a dual face f^* , namely the convex hull of the sites associated to the subset of cells. If no subset of $d + 2$ sites lie on a same sphere, the set of dual faces (of dimensions 0 to d) constitutes a triangulation embedded in \mathbb{R}^d whose vertices are the sites. This triangulation is called the *Delaunay triangulation* of \mathcal{S} , noted $\text{del}(\mathcal{S})$. The correspondence defined above between the faces of $\text{vor}(\mathcal{S})$ and those of $\text{del}(\mathcal{S})$ is a bijection that satisfies: $f \subset g \Rightarrow g^* \subset f^*$. We say that $\text{del}(\mathcal{S})$ is the *geometric dual* of $\text{vor}(\mathcal{S})$. See Figure 12.

A similar construct is known also for power diagrams. Consider the power diagram of a finite set of balls of \mathbb{R}^d . In the same way as for Euclidean Voronoi diagrams, we can associate a triangulation dual to the power diagram of the balls. This triangulation is called the *regular triangulation* of the balls. The vertices of this triangulation are the centers of the balls whose cell is non empty.

We derive two triangulations from Bregman Voronoi diagrams. One has straight edges and captures some important properties of the Delaunay triangulation. However, it is not always the geometric dual of the corresponding Bregman Voronoi diagram. The other one has curved (geodesic) edges and is the geometric dual of the Bregman Voronoi diagram.

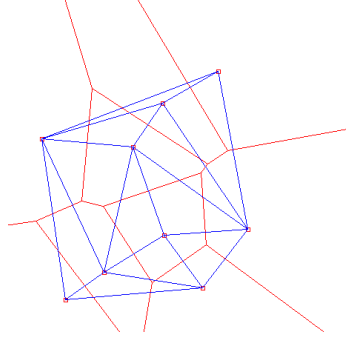


Figure 12: Ordinary Voronoi diagram (red) and geometric dual Delaunay triangulation (blue).

5.1 Bregman Delaunay triangulations

Let $\hat{\mathcal{S}}$ be the lifted image of \mathcal{S} and let \mathcal{T} be the lower convex hull of $\hat{\mathcal{S}}$, i.e. the collection of facets of the convex hull of $\hat{\mathcal{S}}$ whose supporting hyperplanes are below $\hat{\mathcal{S}}$. We assume in this section that \mathcal{S} is in *general position* if there is no subset of $d+2$ points lying on a same Bregman sphere. Equivalently (see Lemma 5), \mathcal{S} is in *general position* if no subset of $d+2$ points $\hat{\mathbf{p}}_i$ lying on a same hyperplane.

Under the general position assumption, each vertex of $\mathcal{H} = \cap_i H_{\hat{\mathbf{p}}_i}^\uparrow$ is the intersection of exactly $d+1$ hyperplanes and the faces of \mathcal{T} are all *simplices*. Moreover the vertical projection of \mathcal{T} is a triangulation $\text{del}_F(\mathcal{S}) = \text{Proj}_{\mathcal{X}}(\mathcal{T})$ of \mathcal{S} embedded in $\mathcal{X} \subseteq \mathbb{R}^d$. Indeed, since the restriction of $\text{Proj}_{\mathcal{X}}$ to \mathcal{T} is bijective, $\text{del}_F(\mathcal{S})$ is a simplicial complex embedded in \mathcal{X} . Moreover, since F is convex, $\text{del}_F(\mathcal{S})$ covers the (Euclidean) convex hull of \mathcal{S} , and the set of vertices of \mathcal{T} consists of all the $\hat{\mathbf{p}}_i$. Consequently, the set of vertices of $\text{del}_F(\mathcal{S})$ is \mathcal{S} . We call $\text{del}_F(\mathcal{S})$ the *Bregman Delaunay triangulation* of \mathcal{S} (see Fig. 13). When $F(\mathbf{x}) = \|\mathbf{x}\|^2$, $\text{del}_F(\mathcal{S})$ is the Delaunay triangulation dual to the Euclidean Voronoi diagram. This duality property holds for symmetric Bregman divergences (via polarity) but not for general Bregman divergences.

We say that a Bregman sphere σ is *empty* if the open ball bounded by σ does not contain any point of \mathcal{S} . The following theorem extends a similar well-known property for Delaunay triangulations whose proof (see, for example [10]) can be extended in a straightforward way to Bregman triangulations using the lifting map introduced in Section 3.2.

Theorem 12 *The first-type Bregman sphere circumscribing any simplex of $\text{del}_F(\mathcal{S})$ is empty. $\text{del}_F(\mathcal{S})$ is the only triangulation of \mathcal{S} with this property when \mathcal{S} is in general position.*

Several other properties of Delaunay triangulations extend to Bregman triangulations. We list some of them.

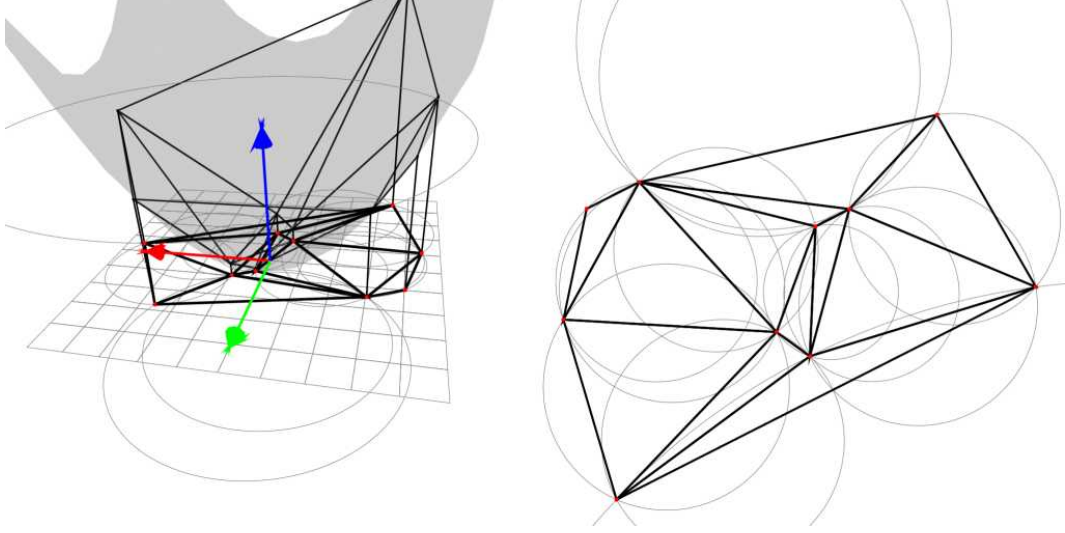


Figure 13: Bregman Delaunay triangulation as the projection of the convex polyhedron \mathcal{T} .

Theorem 13 (Empty ball) *Let ν be a subset of at most $d + 1$ indices in $\{1, \dots, n\}$. The convex hull of the associated points \mathbf{p}_i , $i \in \nu$, is a simplex of the Bregman triangulation of \mathcal{S} iff there exists an empty Bregman sphere σ passing through the \mathbf{p}_i , $i \in \nu$.*

The next property exhibits a local characterization of Bregman triangulations. Let $T(\mathcal{S})$ be a triangulation of \mathcal{S} . We say that a pair of adjacent facets $f_1 = (f, \mathbf{p}_1)$ and $f_2 = (f, \mathbf{p}_2)$ of $T(\mathcal{S})$ is regular iff \mathbf{p}_1 does not belong to the open Bregman ball circumscribing f_2 and \mathbf{p}_2 does not belong to the open Bregman ball circumscribing f_1 (the two statements are equivalent for symmetric Bregman divergences).

Theorem 14 (Locality) *Any triangulation of a given set of points \mathcal{S} (in general position) whose pairs of facets are all regular is the Bregman triangulation of \mathcal{S} .*

Let \mathcal{S} be a given set of points, $\text{del}_F(\mathcal{S})$ its Bregman triangulation, and $\mathcal{T}(\mathcal{S})$ the set of all triangulations of \mathcal{S} . We define the Bregman radius of a d -simplex τ as the radius noted $r(\tau)$ of the smallest Bregman ball containing τ . The following result is an extension of a result due to Rajan for Delaunay triangulations [37].

Theorem 15 (Optimality) *We have $\text{del}_F(\mathcal{S}) = \min_{T \in \mathcal{T}(\mathcal{S})} \max_{\tau \in T} r(\tau)$.*

The proof mimics Rajan's proof [37] for the case of Delaunay triangulations.

5.2 Bregman geodesic triangulations

We have seen in Section 4.3 that the Bregman Voronoi of a set of points \mathcal{S} is the power diagram of a set of balls \mathcal{B}' centered at the points of \mathcal{S}' (Theorem 8). Write $\text{reg}_F(\mathcal{B}')$ for the

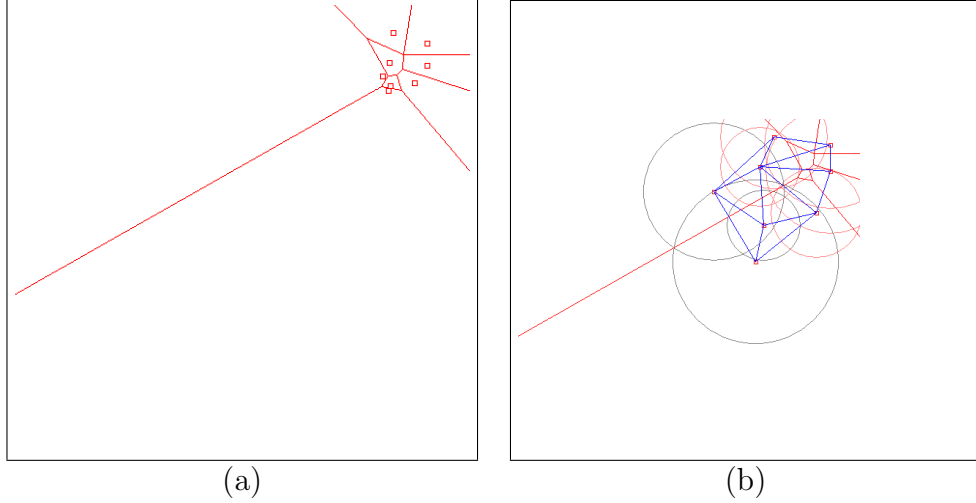


Figure 14: First-type Kullback-Leibler Bregman Voronoi diagram (a) obtained from the corresponding power diagram (b), and its associated dual regular triangulation rooted at gradient vertices (blue).

dual regular triangulation dual to this power diagram. This triangulation⁴ is embedded in \mathcal{X}' and has the points of \mathcal{S}' as its vertices (see Figure 14). The image of this triangulation by $\nabla^{-1}F$ is a *curved triangulation* whose vertices are the points of \mathcal{S} . The edges of this curved triangulation are geodesic arcs joining two sites (see Section 3.3). We call it the *Bregman geodesic triangulation* of \mathcal{S} , noted $\text{del}'_F(\mathcal{S})$ (see Figure 15).

Theorem 16 *The Bregman geodesic triangulation $\text{del}'_F(\mathcal{S})$ is the geometric dual of the 1st-type Bregman Voronoi diagram of \mathcal{S} .*

Proof: We have, noting $\overset{*}{\equiv}$ for the dual mapping, and using Theorem 8

$$\text{vor}_F(\mathcal{S}) \equiv \text{pow}(\mathcal{B}') \overset{*}{\equiv} \text{reg}(\mathcal{B}') = \nabla F(\text{del}'_F(\mathcal{S})).$$

□

Observe that $\text{del}'_F(\mathcal{S})$ is, in general, distinct from $\text{del}_F(\mathcal{S})$, the Bregman Delaunay triangulation introduced in the previous section. However, when the divergence is symmetric, both triangulations are combinatorially equivalent and dual to the Bregman Voronoi diagram of \mathcal{S} . Moreover, they coincide exactly when F is the squared Euclidean distance.

6 Applications

In this section, we give some applications related to computational geometry and machine learning.

⁴Applet at <http://www.csl.sony.co.jp/person/nielsen/BVDapplet/>

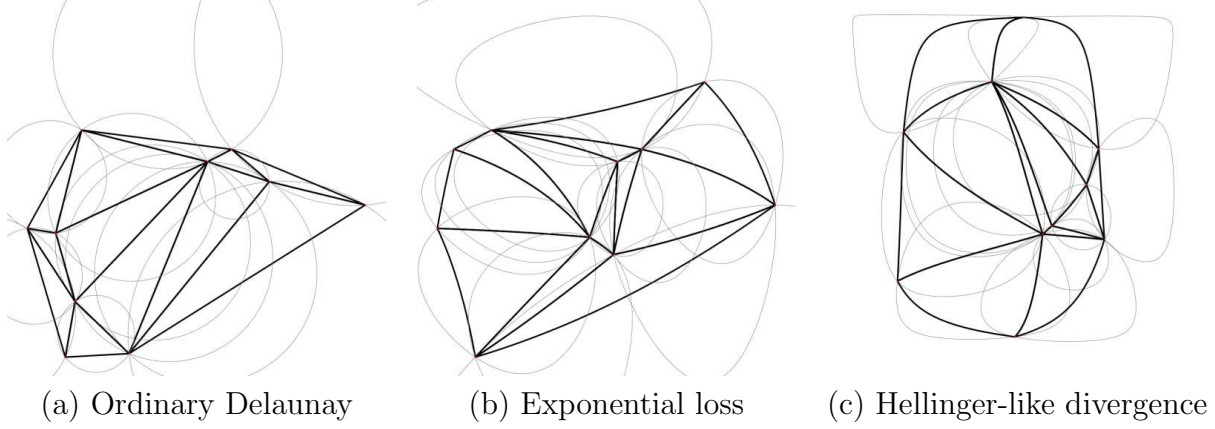


Figure 15: An ordinary Delaunay triangulation (a) and two Bregman geodesic triangulations for the exponential loss (b) and for the Hellinger-like divergence (c).

6.1 Centroidal Bregman Voronoi diagrams and Lloyd quantization

Let \mathcal{D} be a domain of \mathcal{X} and $p(\mathbf{x})$ be a density function defined over \mathcal{D} . We define the *Bregman centroid* of \mathcal{D} as the point $\mathbf{c}^* \in \mathcal{D}$ such that $\mathbf{c}^* = \operatorname{argmin}_{\mathbf{c} \in \mathcal{D}} \int_{\mathbf{x} \in \mathcal{D}} p(\mathbf{x}) D_F(\mathbf{x}||\mathbf{c}) d\mathbf{x}$. The following lemma states that the mass Bregman centroid of \mathcal{D} is *uniquely* defined and *independent* of F .

Lemma 11 *The Bregman centroid of \mathcal{D} coincides with the mass centroid of \mathcal{D} .*

Proof:

$$\begin{aligned}
 \nabla_{\mathbf{c}} \int_{\mathbf{x} \in \mathcal{D}} p(\mathbf{x}) D_F(\mathbf{x}||\mathbf{c}) d\mathbf{x} &= \nabla_{\mathbf{c}} \int_{\mathbf{x} \in \mathcal{D}} p(\mathbf{x}) (F(\mathbf{x}) - F(\mathbf{c}) - \langle \mathbf{x} - \mathbf{c}, \nabla F(\mathbf{c}) \rangle) d\mathbf{x} \\
 &= - \int_{\mathbf{x} \in \mathcal{D}} p(\mathbf{x}) \nabla^2 F(\mathbf{c}) (\mathbf{x} - \mathbf{c}) d\mathbf{x} \\
 &= - \nabla^2 F(\mathbf{c}) \left(\int_{\mathbf{x} \in \mathcal{D}} p(\mathbf{x}) \mathbf{x} d\mathbf{x} - \mathbf{c} \int_{\mathbf{x} \in \mathcal{D}} p(\mathbf{x}) d\mathbf{x} \right).
 \end{aligned}$$

Hence, $\mathbf{c}^* = \frac{\int_{\mathbf{x} \in \mathcal{D}} p(\mathbf{x}) \mathbf{x} d\mathbf{x}}{\int_{\mathbf{x} \in \mathcal{D}} p(\mathbf{x}) d\mathbf{x}}$. \square

When \mathbf{x} is a random variable following the probability density $p(\mathbf{x})$, $\int_{\mathbf{x} \in \mathcal{D}} p(\mathbf{x}) D_F(\mathbf{x}||\mathbf{c}) d\mathbf{x}$ is called the *distortion rate* associated to the representative \mathbf{c} , the optimal distortion-rate function $\int_{\mathbf{x} \in \mathcal{D}} p(\mathbf{x}) D_F(\mathbf{x}||\mathbf{c}^*) d\mathbf{x}$ is called the *Bregman information*, and \mathbf{c}^* is called the *Bregman representative*. The above result states that the optimal distortion rate exists and does not depend on the choice of the Bregman divergence, and that the Bregman representative \mathbf{c}^* is the expectation $E(\mathbf{x})$ of \mathbf{x} . This result extends an analogous result in the discrete case (finite point sets) studied in [6].

Computing a centroidal Bregman Voronoi diagram of k points can be done by means of Lloyd's algorithm [30]. We select an initial set of k points. Then, we iteratively compute a Bregman Voronoi diagram and move the sites to the Bregman centroids of the corresponding cells in the diagram. Upon convergence, the output of the algorithm is a local minimizer of $f((\mathbf{p}_i, V_i), i = 1, \dots, k) = \sum_{i=1}^k \int_{\mathbf{x} \in V_i} D_F(\mathbf{x} || \mathbf{p}_i) d\mathbf{x}$, where $\{\mathbf{p}_i\}_{i=1}^k$ denotes any set of k points of \mathcal{X} and $\{V_i\}_{i=1}^k$ denotes any tessellation of \mathcal{X} into k regions. See [18] for a further discussion and applications of centroidal Voronoi diagrams.

6.2 ε -nets

Lloyd's algorithm intends to find a best set of k points for a given k so as to minimize a least-square criterion. Differently, we may want to sample a compact domain $\mathcal{D} \subset \mathcal{X}$ up to a given precision while minimizing the number of samples. Instead of a least-square criterion, we define the error associated to a sample P as $\text{error}(P) = \max_{\mathbf{x} \in \mathcal{D}} \min_{\mathbf{p}_i \in P} D_F(\mathbf{x} || \mathbf{p}_i)$. A finite set of points P of \mathcal{D} is an ε -sample of \mathcal{D} iff $\text{error}(P) \leq \varepsilon$.

An ε -sample P is called an ε -net if it satisfies the sparsity condition: $\max(D_F(\mathbf{p} || \mathbf{q}), D_F(\mathbf{q} || \mathbf{p})) > \varepsilon$ for any two points \mathbf{p} and \mathbf{q} in P .

We will see how to construct an ε -net. For simplicity, we assume in the rest of the section that \mathcal{D} is a convex polytope. Extending the results to more general domains is possible.

Let $P \subset \mathcal{D}$, $\text{vor}_F(P)$ be the Bregman Voronoi diagram of P and $\text{vor}_{F|\mathcal{D}}(P)$ be its restriction to \mathcal{D} . Write V for the set of vertices of $\text{vor}_{F|\mathcal{D}}(P)$. V consists of vertices of $\text{vor}_F(P)$ and intersection points between the edges of $\text{vor}_F(P)$ and the boundary of \mathcal{D} . The following lemma states that $\text{error}(P)$ can be computed by examining only a finite number of points, namely the points of V .

Lemma 12 $\text{error}(P) = \max_{\mathbf{v} \in V} \min_{\mathbf{p}_i \in P} D_F(\mathbf{x} || \mathbf{p}_i)$.

Proof: Let $\mathbf{x} \in \mathcal{D}$, $\mathbf{p}_{\mathbf{x}}$ the point of P closest to \mathbf{x} and $V_{\mathbf{x}}$ the associated cell of $\text{vor}_{F|\mathcal{D}}(P)$ (which contains \mathbf{x}). $V_{\mathbf{x}}$ is a bounded polytope whose vertices belongs to V . Let \mathbf{w} be the vertex of $V_{\mathbf{x}}$ most distant from $\mathbf{p}_{\mathbf{x}}$. We have $D_F(\mathbf{x} || \mathbf{p}_{\mathbf{x}}) \leq D_F(\mathbf{w} || \mathbf{p}_{\mathbf{x}})$. This is a consequence of the convexity of F and of the fact that $D_F(\mathbf{x} || \mathbf{p})$ is measured by the vertical distance between $\hat{\mathbf{x}}$ and $H_{\mathbf{p}}$ (Lemma 1). \square

An ε -net of \mathcal{D} can be constructed by the following greedy algorithm originally proposed by Ruppert in the context of mesh generation [39]. See also [20]. We initialize the sample set P_0 with d points of \mathcal{D} lying at distance greater than ε from one another. Then, at each step, the algorithm looks for the point \mathbf{v}_i of \mathcal{D} that is the furthest (for the considered Bregman divergence) from the current set of samples P_i . By Lemma 12, this step reduces to looking at the vertices of $\text{vor}_{F|\mathcal{D}}(P_i)$. If $D_F(\mathbf{x} || \mathbf{v}_i) \leq \varepsilon$, the algorithm stops. Otherwise, we take \mathbf{v}_i as a new sample point, i.e. $\mathbf{p}_{i+1} = \mathbf{v}_i$, we update the set of sample points, i.e. $P_{i+1} = P_i \cup \{\mathbf{p}_{i+1}\}$, and insert \mathbf{p}_{i+1} in the Bregman Voronoi diagram of the sample points. Upon termination, the

set of sample points P_t satisfies the hypothesis of Lemma 12 and therefore P_t is an ε -sample of \mathcal{D} . Moreover, for any two points \mathbf{p} and \mathbf{q} of P_t , we have $D_F(\mathbf{p}||\mathbf{q}) > \varepsilon$ or $D_F(\mathbf{q}||\mathbf{p}) > \varepsilon$, depending on whether \mathbf{p} has been inserted after or before \mathbf{q} . Indeed, we only insert a point if its divergence to the points of the current sample is greater than ε . Hence, P_t is an ε -net of \mathcal{D} .

To prove that the algorithm terminates, we need the following lemma. Given a Bregman ball $B(\mathbf{c}, r)$, we define the biggest Euclidean ball $EB(\mathbf{c}, r')$ contained in $B(\mathbf{c}, r)$ and the smallest Euclidean ball $EB(\mathbf{c}, r'')$ containing $B(\mathbf{c}, r)$.

Lemma 13 *Let F be a strictly convex function of class C^2 , there are constants γ' and γ'' (that do not depend on \mathbf{c} nor on r) such that $r'^2 \geq \gamma' r$ and $r''^2 \leq \gamma'' r$.*

Proof: According to Taylor's formula, there exists a point \mathbf{t} of the open segment $\mathbf{x}\mathbf{c}$ such that

$$F(\mathbf{x}) = F(\mathbf{c}) + \langle \mathbf{x} - \mathbf{c}, \nabla F(\mathbf{c}) \rangle + \frac{1}{2} (\mathbf{x} - \mathbf{c})^T \nabla^2 F(\mathbf{t}) (\mathbf{x} - \mathbf{c}).$$

Hence,

$$D_F(\mathbf{x}||\mathbf{c}) = F(\mathbf{x}) - F(\mathbf{c}) - \langle \mathbf{x} - \mathbf{c}, \mathbf{c}' \rangle = \frac{1}{2} (\mathbf{x} - \mathbf{c})^T \nabla^2 F(\mathbf{t}) (\mathbf{x} - \mathbf{c}), \quad (14)$$

where \mathbf{t} is a point of the open segment $\mathbf{x}\mathbf{c}$.

Since F is strictly convex, the Hessian matrix is positive definite (i.e., $\mathbf{x}^T \nabla^2 F(\mathbf{t}) \mathbf{x} > 0$ for all \mathbf{x} in \mathcal{X}), and the domain \mathcal{D} being compact, there exist two constants η' and η'' such that, for any $\mathbf{y} \in \mathcal{D}$, $0 < \eta'' \leq \|\nabla^2 F(\mathbf{y})\| \leq \eta'$. If $\|\mathbf{x} - \mathbf{c}\|^2 > \frac{2r}{\eta''}$ (Fröbenius matrix norm), we deduce from Equation (14) that $D_F(\mathbf{x}||\mathbf{c}) > r$. Therefore, $B(\mathbf{c}, r) \subset EB(\mathbf{c}, \sqrt{\frac{2r}{\eta''}})$.

If $\|\mathbf{x} - \mathbf{c}\|^2 \leq \frac{2r}{\eta'}$, we have using again Equation (14)

$$D_F(\mathbf{x}||\mathbf{c}) \leq \frac{\eta'}{2} \|\mathbf{x} - \mathbf{c}\|^2 \leq r.$$

Therefore, $EB(\mathbf{c}, \sqrt{\frac{2r}{\eta'}}) \subset B(\mathbf{c}, r)$. \square

Let \mathbf{p} and \mathbf{q} be two points such that $D_F(\mathbf{p}||\mathbf{q}) = r$. Observing that $EB(\mathbf{p}, r') \subseteq EB(\mathbf{p}, \|\mathbf{p} - \mathbf{q}\|) \subseteq EB(\mathbf{p}, r'')$, we deduce from the above lemma that

$$\sqrt{\gamma' r} \leq \|\mathbf{p} - \mathbf{q}\| \leq \sqrt{\gamma'' r} \quad (15)$$

and

$$\frac{\gamma'}{\gamma''} D_F(\mathbf{p}||\mathbf{q}) \leq D_F(\mathbf{q}||\mathbf{p}) \leq \frac{\gamma''}{\gamma'} D_F(\mathbf{p}||\mathbf{q}).$$

Another consequence of the lemma is that the volume of any Bregman ball of radius at least $r > 0$, is bounded away from 0 (when F is of class C^2). Hence, since \mathcal{D} is compact, the

algorithm cannot insert infinitely many points and therefore terminates. Moreover, the size of the sample output by the algorithm can be bounded, as stated in the next lemma. Write $\mathcal{D}^{\leq \varepsilon} = \{\mathbf{x} \mid \exists \mathbf{y} \in \mathcal{D}, \|\mathbf{x} - \mathbf{y}\| \leq \varepsilon\}$.

Lemma 14 *If F is of class C^2 , the algorithm terminates. If P_t denotes the final set of sample points, we have $|P_t| = O\left(\frac{\text{vol}(\mathcal{D})}{\varepsilon^{d/2}}\right)$.*

Proof: We have already shown that the algorithm terminates. Let P_t be the set of points that have been inserted by the algorithm, excluding the initial set (of constant size). Let $\tau(\mathbf{x}) = \inf\{r : |EB(\mathbf{x}, r) \cap P_t| \geq 2\}$ and $B_p = EB(\mathbf{p}, \frac{\tau(\mathbf{p})}{2})$, $\mathbf{p} \in P_t$. It is easy to see that τ is 1-Lipschitz and that the Euclidean balls B_p , $\mathbf{p} \in P_t$ are disjoint. Let \mathbf{q} be a point of P_t closest to \mathbf{p} : $\tau(\mathbf{p}) = \|\mathbf{p} - \mathbf{q}\|$ and, as noticed above, $\max(D_F(\mathbf{p}||\mathbf{q}), D_F(\mathbf{q}||\mathbf{p})) > \varepsilon$. Eq. 15 then implies that $\tau(\mathbf{p}) = \|\mathbf{p} - \mathbf{q}\| \geq \sqrt{\gamma'} \varepsilon$. Consider now the midpoint \mathbf{m} of $\mathbf{p}\mathbf{q}$ and write \mathbf{t} for the point of P_t that minimizes $D_F(\mathbf{m}||\cdot)$. Since \mathcal{D} is convex, $\mathbf{m} \in \mathcal{D}$ and, according to the definition of \mathbf{q} , $\|\mathbf{m} - \mathbf{p}\| \leq \|\mathbf{m} - \mathbf{t}\|$. Eq. 15 and the fact that P_t is an ε -sample of \mathcal{D} then yield $\|\mathbf{m} - \mathbf{t}\| \leq \sqrt{\gamma''} \varepsilon$. In summary, we have

$$\sqrt{\gamma'} \varepsilon \leq \tau(\mathbf{p}) = \|\mathbf{p} - \mathbf{q}\| \leq 2\sqrt{\gamma''} \varepsilon. \quad (16)$$

The right inequality shows that all the balls B_p , $\mathbf{p} \in P_t$, are contained in $\mathcal{D}^{\leq \eta}$ where $\eta = \sqrt{\gamma''} \varepsilon$. We can now bound the size of P_t .

$$\begin{aligned} \int_{\mathcal{D}^{\leq \eta}} \frac{d\mathbf{x}}{\tau^d(\mathbf{x})} &\geq \sum_{\mathbf{p} \in P_t} \int_{B_p \cap \mathcal{D}^{\leq \eta}} \frac{d\mathbf{x}}{\tau^d(\mathbf{x})} \quad (\text{the balls } B_p \text{ have disjoint interiors}) \\ &\geq \sum_{\mathbf{p} \in P} \frac{\text{vol}(B_p \cap \mathcal{D}^{\leq \eta})}{(\frac{3}{2}\tau(\mathbf{p}))^d} \quad (\tau(\mathbf{x}) \leq \tau(\mathbf{p}) + \|\mathbf{p} - \mathbf{x}\| \leq \frac{3}{2}\tau(\mathbf{p})) \\ &\geq \frac{C}{3^d} |P_t| \end{aligned}$$

where $C = \frac{\pi^p}{p!}$ if $d = 2p$ and $C = \frac{2^{2p-1}(p-1)! \pi^{p-1}}{(2p-1)!}$ if $d = 2p - 1$.

Using again the Lipschitz property of τ and Eq 16, we have for all $\mathbf{x} \in B_p$

$$\tau(\mathbf{x}) \geq \tau(\mathbf{p}) - \|\mathbf{x} - \mathbf{p}\| \geq \frac{1}{2} \tau(\mathbf{p}) \geq \frac{1}{2} \sqrt{\gamma'} \varepsilon$$

We deduce

$$|P_t| \leq \left(\frac{6}{\sqrt{\gamma'}}\right)^d \frac{1}{C \varepsilon^{d/2}} \int_{\mathcal{D}^{\leq \eta}} d\mathbf{x}.$$

□

A geometric object O is said α -fat [7] if the ratio $\frac{r^+}{r^-}$ of the radius r^+ of the smallest ball enclosing O over the radius r^- of the largest ball inscribed in O is bounded by α : $\frac{r^+}{r^-} \leq \alpha$.

Euclidean balls are therefore 1-fat, namely the fattest objects. It has been shown that considering the fatness factor for a set of objects yields in practice efficient tailored data-sensitive algorithms [7] by avoiding bad configurations of sets of skinny objects. A direct consequence of Lemma 13 is that Bregman balls (in fixed dimensions) are fat (i.e., $\alpha = O(1)$) on any *compact* domain:

Corollary 3 *For C^2 Bregman generator functions, Bregman balls on any compact domain are fat.*

Proof: Indeed, consider any Bregman ball defined on a compact domain for a C^2 strictly convex and differentiable Bregman generator function F . Its fatness α is upper bounded by $\sqrt{\frac{\gamma'}{\gamma''}}$, where γ' and γ'' are the two constants (depending on F and \mathcal{D}) of Lemma 13. Recall that Lemma 13 considers concentric Euclidean balls ham sandwiching a Bregman ball, all centered at position \mathbf{c} . We have $\alpha \leq \frac{r^+}{r^-} \leq \frac{r^+}{r_c^-} \leq \frac{r_c^+}{r_c^-} = O(1)$ since $r_c^- \leq r^-$ and $r_c^+ \geq r^+$, where r_c^+ (respectively, r_c^-) denote the radius of the smallest enclosing (respectively, largest inscribed) Euclidean ball centered at \mathbf{c} . The fatness property simply means that we can cover any Bregman ball by a constant number of (convex) Euclidean balls. \square

Thus, since Bregman balls are fat on compact domains, we can build efficient data-structures for point location with applications to piercing (geometric 0-transversal) and others, as described in [19].

6.3 VC-dimension, classification and learning

Some important classification rules rely on Voronoi diagrams; furthermore, the analysis of classification rules (complexity or statistical generalization) sometimes makes use of concepts closely related to Voronoi diagrams. Extending the rules and analyses to arbitrary Bregman divergences, with important related consequences (such as the eventual lost of convexity) is thus particularly interesting for classification, and we review here some notable consequences.

In supervised classification, we are generally interested in capturing the joint structure of \mathcal{X} and a set of *classes*, $\{0, 1\}$ in the simplest case. For this objective, we build *representations of concepts*, i.e. functions that map \mathcal{X} to the set of classes. A concept class \mathcal{H} is a set of concept representations $h : \mathcal{X} \rightarrow \{0, 1\}$; for example, should h be a Bregman ball, it would classify 0 the points outside the ball, and 1 the points inside. Armed with these definitions, our supervised classification problem becomes the following one. A so-called *target* concept, c , which is unknown, labels the points of \mathcal{X} ; we have access to its labeling throughout a sampling process: we retrieve *examples* (i.e., pairs $(\mathbf{x}, c(\mathbf{x}))$), independently at random, according to some *unknown* but *fixed* distribution \mathcal{D} over the set $\{(\mathbf{x}, c(\mathbf{x})) : \mathbf{x} \in \mathcal{X}\}$. The question is: what are the conditions on \mathcal{H} that guarantee the possibility to build, within reasonable time, some $h \in \mathcal{H}$ agreeing as best as possible with c , with high probability? While the complexity requirement is usual in computer science, the fact that we require

adequacy with high probability better than systematically is also a necessary requirement, as there is always the possibility of an extremely bad sampling that would prevent any efficient learning (*e.g.* we have drawn the same example all the time). In general, rather than directly sampling the domain, we work with a finite data set \mathcal{S} of examples which is supposed to be sampled this way.

From the statistical standpoint, learning requires to find a good balance between the accuracy, *i.e.* the goodness-of-fit of h as measured on \mathcal{S} , and the *capacity* of \mathcal{H} , *i.e.* its ability to *learn* (or fit in generalization) the data with the smallest number of errors. Consider for example geometric figures in the plane and the “square” concept. Intuitively, an \mathcal{H} with too large capacity is like the person who picks a huge quantity of geometric figures including squares, memorizes each of them, and then rejects every square that would not exactly be in its collection (edge lengths, colors, etc.). An \mathcal{H} with too little capacity is like the lazy person who keeps as sole concept the fact that squares have four edges. Both extremal situations mean little generalization capabilities, but for different reasons.

There have been intensive lines of works on the measures of this capacity, and one of the most popular is the VC-dimension [17]. Informally, the VC-dimension of \mathcal{H} is the size of the largest dataset \mathcal{S} for which \mathcal{H} *shatters* \mathcal{S} , *i.e.* for which \mathcal{H} contains all the classifiers that could perform any of the $2^{|\mathcal{S}|}$ possible labelings of the data. To be more formal, let $\Pi_{\mathcal{H}}(\mathcal{S}) = \{(h(\mathbf{p}_1), h(\mathbf{p}_2), \dots, h(\mathbf{p}_n)) \mid h \in \mathcal{H}\}$ denote the set of all distinct tuples of labels on \mathcal{S} that can be performed by elements of \mathcal{H} . While it always holds that $\Pi_{\mathcal{H}}(n) \leq 2^n$, the maximal n for which $\Pi_{\mathcal{H}}(n) = 2^n$ is the VC-dimension of \mathcal{H} , $\text{VCDIM}(\mathcal{H})$. The importance of the VC-dimension comes from the fact that it allows to bound the behavior of the empirical optimal classifier in a distribution-free manner [17]. In particular, if the VC-dimension is finite, the average error probability of the empirical optimal classifier tends to 0 when the size of the training data set increases. The following lemma proves that the VC-dimension of Bregman balls is the same as for linear separators, and this does not depend on the choice of F .

Theorem 17 *The VC dimension of the class of all Bregman balls B_F of \mathbb{R}^d (for any given strictly convex and differentiable function F) is $d + 1$.*

Proof: We use the lifting map introduced in Section 3.2. Given a set \mathcal{S} of points in \mathbb{R}^d , we lift them onto \mathcal{F} , obtaining $\hat{\mathcal{S}} \in \mathbb{R}^{d+1}$.

Let B_F be a Bregman ball and write σ for the Bregman sphere bounding B_F . From Lemma 5, we know that, for any $\mathbf{p} \in \mathbb{R}^d$, $\mathbf{p} \in B$ iff $\hat{\mathbf{p}} \in H_{\sigma}^{\downarrow}$. For a given function F , let \mathcal{B}_F denote the set of all Bregman balls, and let \mathcal{H}_F denote the set of all lower halfspaces of \mathbb{R}^{d+1} . It follows from the observation above that \mathcal{B} shatters \mathcal{S} iff \mathcal{H} shatters $\hat{\mathcal{S}}$. Hence the VC dimension of \mathcal{B} over the sets of points of \mathbb{R}^d is equal to the VC dimension of \mathcal{H} over the sets of points of $\mathcal{F} \subset \mathbb{R}^{d+1}$.

Since the points of $\hat{\mathcal{S}}$ are in convex position, they are shattered by \mathcal{H} iff the affine hull of their convex hull is of dimension strictly less than the dimension of the embedding space, *i.e.*

$d + 1$, which happens iff $|\mathcal{S}| < d + 2$. Indeed otherwise, the subset of vertices of any facet of the upper convex hull of $\hat{\mathcal{S}}$ cannot be obtained by intersecting $\hat{\mathcal{S}}$ with a lower halfspace (an upper halfspace would be required). Hence, the VC dimension of Bregman balls is at most $d + 1$.

It is exactly $d + 1$ since any set of $d + 1$ points on \mathcal{F} in general position generates a d -dimensional affine hull \mathcal{A} that cannot be shattered by less than $d + 1$ hyperplanes of \mathcal{A} . The same result plainly holds for hyperplanes of \mathbb{R}^{d+1} since we can associate to each hyperplane h of \mathcal{A} a hyperplane H of \mathbb{R}^{d+1} such that $h = H \cap \mathcal{A}$. \square

This result does not fall into the general family of VC bounds for concept classes parameterized by polynomial-based predicates [23], it is mostly exact, and it happens not to depend on the choice of the Bregman divergence. This has a direct consequence for classification, which is all the more important as Bregman balls are not necessarily convex (see Figure 5). Because the capacity of Bregman balls is not affected by the divergence, if we fit this divergence in order to minimize the empirical risk (risk estimated on \mathcal{S}), then there is an efficient minimization of the true risk (risk estimated on the full domain \mathcal{X}), as well. There is thus little impact (if any) on overfitting, one important pitfall for classification, usually caused by over-capacitating the classifiers by tuning too many parameters.

Some applications of our results in supervised learning also meet one of the oldest classification rule: the k -Nearest Neighbors (k -NN) rule [22], in which a new observation receives the majority class among the set of its k nearest neighbors, using *e.g.* k -order Voronoi diagrams of \mathcal{S} (Section 4.4). Various results establish upperbounds for the k -NN rule that depend on the Bayes risk (the true risk of the best possible rule) [17]. The choice of the proximity notion between observations (it is often not a metric for complex domains) is crucial: if it is too simple or oversimplified, it degrades the k -NN results and may even degrade Bayes risk as well; if it is too complicated or complexified, it may degrade the test results via the capacity of the rule. Searching for accurate “distance” notions has been an active field of research in machine learning in the past decade [42]. Our results on the linearity of the Bregman Voronoi diagrams essentially show that we can mix arbitrary Bregman divergences for heterogenous data (mixing binary, real, integer values, etc.) without losing anything from the capacity standpoint.

Range spaces of finite VC-dimensions have found numerous applications in Combinatorial and Computational Geometry. We refer to Chazelle’s book for an introduction to the subject and references wherein [15]. In particular, Brönnimann and Goodrich [13] have proposed an almost optimal solution to the disk cover algorithm, i.e. to find a minimum number of disks in a given family that cover a given set of points. Theorem 17 allows to extend this result to arbitrary Bregman ball cover (see also [21]).

7 Conclusion

We have defined the notion of Bregman Voronoi diagrams and showed how these geometric structures are a natural extension of ordinary Voronoi diagrams. Bregman Voronoi diagrams share with their Euclidean analogs surprisingly similar combinatorial and geometric properties. We hope that our results will make Voronoi diagrams and their relatives applicable in new application areas. In particular, Bregman Voronoi diagrams based on various entropic divergences are expected to find applications in information retrieval (IR), data mining, knowledge discovery in databases, image processing (e.g., see [24]). The study of Bregman Voronoi diagrams raises the question of revisiting computational geometry problems in this new light. This may also allow one to tackle uncertainty ('noise') in computational geometry for fundamental problems such as surface reconstruction or pattern matching.

A limitation of Bregman Voronoi diagrams is their combinatorial complexity that depends exponentially on the dimension. Since many applications are in high dimensional spaces, building efficient data-structures is a major avenue for further research.

Acknowledgements

Frédéric Chazal, David Cohen-Steiner and Mariette Yvinec are gratefully acknowledged for their comments on this paper. The work by the second author has been partially supported by the project GeoTopAl (1555) of the Agence Nationale de la Recherche (ANR).

References

- [1] S. Amari and H. Nagaoka. *Methods of Information Geometry*. Oxford University Press, ISBN 0-8218-0531-2, 2000.
- [2] A. Ben-Hur, D. Horn, H. T. Siegelmann, and V. Vapnik. Support Vector Clustering. *Journal of Machine Learning Research*, (2):125-137, 2001.
- [3] F. Aurenhammer. Power diagrams: Properties, algorithms and applications. *SIAM Journal of Computing*, 16(1):78-96, 1987.
- [4] F. Aurenhammer and H. Imai. Geometric relations among voronoi diagrams. In *4th Annual Symposium on Theoretical Aspects of Computer Sciences (STACS)*, pp. 53-65, 1987.
- [5] F. Aurenhammer and R. Klein. Voronoi Diagrams. In J. Sack and G. Urrutia (Eds), *Handbook of Computational Geometry, Chapter V*, pp. 201-290. Elsevier Science Publishing, 2000.

- [6] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research (JMLR)*, 6:1705–1749, 2005.
- [7] M. de Berg, M. Katz, F. van der Stappen, and J. Vleugels. Realistic input models for geometric algorithms. *Algorithmica* 34:81–97, 2002.
- [8] J.-D. Boissonnat and M. Karavelas. On the combinatorial complexity of Euclidean Voronoi cells and convex hulls of d -dimensional spheres. In *Proc. 14th ACM-SIAM Sympos. Discrete Algorithms (SODA)*, pp. 305–312, 2003.
- [9] J.-D. Boissonnat, C. Wormser, and M. Yvinec. Anisotropic diagrams: Labelle Shewchuk approach revisited. In *17th Canadian Conference on Computational Geometry (CCCG)*, pp. 266–269, 2005.
- [10] J.-D. Boissonnat and M. Yvinec. *Algorithmic Geometry*. Cambridge University Press, New York, NY, USA, 1998.
- [11] J.-D. Boissonnat, C. Wormser, and M. Yvinec. Curved Voronoi diagrams. In J.-D. Boissonnat and M. Teillaud (Eds) *Effective Computational Geometry for Curves and Surfaces*, pp. 67–116. Springer-Verlag, Mathematics and Visualization, 2007.
- [12] L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7:200–217, 1967.
- [13] H. Brönnimann and M. T. Goodrich. Optimal set covers in finite VC-dimension. *Discrete & Computational Geometry*, 14(4):463–479, 1995.
- [14] B. Chazelle. An optimal convex hull algorithm in any fixed dimension. *Discrete Computational Geometry*, 10:377–409, 1993.
- [15] B. Chazelle. *The Discrepancy Method*. Cambridge University Press, Cambridge, U.K., 2000.
- [16] I. Csiszár. Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. *Ann. Stat.*, 19:2032–2066, 1991.
- [17] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.
- [18] Q. Du, V. Faber, and M. Gunzburger. Centroidal Voronoi tessellations: Applications and algorithms. *SIAM Review*, 41:637–676, 1999.
- [19] A. Efrat, M. J. Katz, F. Nielsen, and M. Sharir. Dynamic data structures for fat objects and their applications. *Comput. Geom. Theory Appl.*, 15(4):215–227, 2000.
- [20] Y. Eldar, M. Lindenbaum, M. Porat, and Y. Y. Zeevi. The farthest point strategy for progressive image sampling. *IEEE Trans. on Image Processing*, 6(9):1305–1315, 1997.

- [21] G. Even, D. Rawitz, and S. Shahar. Hitting sets when the VC-dimension is small. *Inf. Process. Lett.*, 95(2):358–362, 2005.
- [22] E. Fix and J. L. Hodges. Discriminatory analysis, nonparametric discrimination. Technical Report TR-21-49-004, Rept 4, USAF School of Aviation Medicine, Randolph Field, TX, 1951.
- [23] P.-W. Goldberg and M. Jerrum. Bounding the Vapnik-Chervonenkis dimension of concept classes parameterized by real numbers. *Machine Learning*, 18:131–148, 1995.
- [24] M. Inaba and H. Imai. Geometric clustering models for multimedia databases. In *Proceedings of the 10th Canadian Conference on Computational Geometry (CCCG’98)*, 1998.
- [25] M. Inaba and H. Imai. Geometric clustering for multiplicative mixtures of distributions in exponential families. In *Proceedings of the 12th Canadian Conference on Computational Geometry (CCCG’00)*, 2000.
- [26] R. Klein. *Concrete and Abstract Voronoi Diagrams*, volume 400 of *Lecture Notes in Computer Science*. Springer, 1989. ISBN 3-540-52055-4.
- [27] F. Labelle and J. R. Shewchuk. Anisotropic voronoi diagrams and guaranteed-quality anisotropic mesh generation. In *Proc. 19th Symposium on Computational Geometry (SoCG)*, pages 191–200, New York, NY, USA, 2003. ACM Press.
- [28] J. Lafferty. Additive models, boosting, and inference for generalized divergences. In *Proc. 12th Conference on Computational learning theory*, 125-133, 1999.
- [29] D.-D. Le and S. Satoh. Ent-Boost: Boosting Using Entropy Measure for Robust Object Detection. In *Proc. 18th International Conference on Pattern Recognition*, pp. 602-605, 2006.
- [30] S. P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–136, 1982.
- [31] P. McMullen. The maximum numbers of faces of a convex polytope. *J. Combinatorial Theory, Ser. B*, 10:179–184, 1971.
- [32] F. Nielsen. *Visual Computing: Geometry, Graphics, and Vision*. Charles River Media/Thomson Delmar Learning, ISBN 1584504277, 2005.
- [33] M. Teillaud O. Devillers, S. Meiser. The space of spheres, a geometric tool to unify duality results on voronoi diagrams. Technical Report No.1620, INRIA, 1992.
- [34] K. Onishi and H. Imai. Voronoi diagram in statistical parametric space by Kullback-Leibler divergence. In *Proc. 13th Symposium on Computational Geometry (SoCG)*, pages 463–465, New York, NY, USA, 1997. ACM Press.

- [35] K. Onishi and H. Imai. Voronoi diagrams for an exponential family of probability distributions in information geometry. In *Japan-Korea Joint Workshop on Algorithms and Computation*, 1997.
- [36] S. Pion and M. Teillaud. 3d triangulation data structure. In CGAL Editorial Board, editor, *CGAL-3.2 User and Reference Manual*. 2006.
- [37] V. T. Rajan. Optimality of the Delaunay triangulation in \mathbb{R}^d . *Discrete & Computational Geometry*, 12:189–202, 1994.
- [38] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, New Jersey, 1970.
- [39] J. Ruppert. A Delaunay refinement algorithm for quality 2-dimensional mesh generation. *J. Algorithms*, 18:548–585, 1995.
- [40] K. Sadakane, H. Imai, K. Onishi, M. Inaba, F. Takeuchi, and K. Imai. Voronoi diagrams by divergences with additive weights. In *Proc. 14th Symposium on Computational Geometry (SoCG)*, pages 403–404, New York, NY, USA, 1998. ACM Press.
- [41] M. Sharir. Almost tight upper bounds for lower envelopes in higher dimensions. *Discrete Comput. Geom.*, 12:327–345, 1994.
- [42] D. Randall Wilson and Tony R. Martinez. Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research*, 1:1–34, 1997.